Characterization of wild trypanosome coats using spliced-leader sequence enrichment and RNA-Seq

by

Jaime So

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland

April 2019

Abstract:

African Trypanosomes defend themselves against host antibody in the bloodstream and tissue through antigenic variation of a highly immunogenic coat of variant surface glycoproteins (VSG). These parasites persist in the host by utilizing a large genomic repertoire of VSG genes and pseudogenes to switch the expression of their surface coats throughout infection. Much of the current knowledge regarding antigenic variation in African trypanosomes is based on studies of VSG expression in lab-adapted strains of *Trypanosoma brucei*. The study of gene expression in wild trypanosome populations is particularly difficult due to the low parasitemia of natural infections; RNA isolated from whole blood or tissue contains a very small fraction of trypanosome RNA in a mixture of host RNA. To overcome this problem, we developed an RNA-seq library preparation protocol which depletes host ribosomal RNA and messenger RNAs. This protocol takes advantage of the 5' spliced-leader (SL) sequence present on mature messenger RNA (mRNA)s in all trypanosomatid species. Using a biotin-streptavidin pulldown technique, we achieve specific enrichment of parasite transcripts from whole host blood RNA samples. Our parasite-enriched transcriptomes have the potential to allow us to characterize the VSG expression profile and identify other highly expressed putative surface proteins of African trypanosomes across geographically and temporally distinct natural infections. This SL enrichment technique can easily be adapted and applied to *in vivo* studies for any the African trypanosome species by using each of their unique SL sequences as biotinylated baits.

Primary Reader: Dr. Monica Mugnier

Secondary Reader: Dr. Janet Markle

Acknowledgements:

I am very grateful to my advisor, Monica Mugnier, who has had faith in my abilities as a scientist since I first met her. Her mentorship has taught me so much about creative problem solving and finding the value in failed experiments. I feel I am well prepared for moving forward and working towards earning a Ph.D. because her constant support has really given me the confidence to pursue all kinds of things that I was unsure of before. She is an incredible teacher and mentor who brings out the strengths in all of her students.

I would like to thank my reader, Janet Markle, for her time and thoughtful advice. Her feedback and questions helped me to give my thesis context and encouraged me to think of the broader scope of my research.

My time in the Mugnier lab would not have been nearly as enjoyable if not for Bryce Bobb, Gabriela Romero, Alex Beaver, Jaclyn Smith, and Lucy Zhang. They are all excellent and very distracting conversationalists, but they're also great help for brainstorming experiments and analysis. The struggle to figure out how to use R studio is much better with good company.

Thanks to our collaborators at the University of Ghana, Dr. Theresa Manful Gwira and her students, for diagnosing trypanosomiasis in cattle and collecting the blood RNA to send to us for this experiment.

I'd also like to thank the wonderful people at the sequencing core facilities at JHU. David Mohr at the GRCF for making sure our libraries and sequencing machines are

optimized to get best results. Anne Jedlicka and Amanda Dziedzic for running the TapeStation analysis after all six of my library clean-ups. I'm sure they were as happy as I was when the dimers were finally gone.

Finally, I appreciate my family and William, for supporting me though everything I do. They always push me to do my best and believe in me even when I don't.

Table of Contents

Abstract: ..... *ii*

Acknowledgements: ..... *iii*

List of Tables ..... *vi*

List of Figures ..... *vi*

Chapter 1) Introduction *1*

1.1) General epidemiology and disease burden imposed by African Trypanosomes **1**

1.2) Antigenic Variation in African Trypanosomes **3**

1.3) Outstanding Questions **6**

Chapter 2) Developing a method for performing RNA-seq transcriptome analysis on field isolates of African trypanosomes *8*

2.1) Introduction and rationale **8**

2.2) Spliced-leader pulldown **12**

2.3) Validating the method: comparing oligo d(T) and SL selection **15**

2.4) Optimizing SL enrichment **24**

2.5) Application of SL pulldown on RNA from infected cow blood **29**

Chapter 3) Discussion *33*

Chapter 4) Methods *39*

4.1) *T. brucei* cell culture and RNA extraction **39**

4.2) Mouse RNA controls **40**

4.3) Mag-bind bead clean-ups **40**

4.4) SL and oligo d(T) library preparations for validation **41**

Testing Enrichment: 41

Testing Gene Expression: 44

4.5) Enrichment optimization **45**

4.6) SL-sequence enrichment and RNA-seq **49**

References *53*

Curriculum Vitae: *60*

List of Tables

Table 1) A list of spliced-leader sequences for various Trypansosomatid species

Table 2) The gene ID, annotation, and fold-change difference of all genes found to be overrepresented in RNA-seq libraries prepared after SL-sequence enrichment

Table 3) Outline of all oligo hybridization conditions tested during optimization of the biotin-streptavidin SL-sequence pull-down

List of Figures

Figure 1) Hypothetical model of the VSG-antibody interaction

Figure 2) Diagram of cis and trans-splicing mRNA processing mechanisms

Figure 3) Proposed workflow for enriching trypanosome mRNA transcripts from infected animal blood samples

Figure 4) Plot of enrichment of reads aligning to trypanosome reference genome achieved by SL pull-down and RNA-seq

Figure 5) Plots of gene expression RPKMs calculated for oligo d(T) and SL-sequence selected libraries with MULTo for either coding sequences or exons

Figure 6) Plots of the total number of reads processed, percentage of reads aligned to the full reference, percentage of reads aligned to mRNA, and percentage of reads aligned to rRNA by library selection type

Figure 7) Example RT-qPCR standard curves used to estimate the abundance of trypanosome and mouse material the experimental pull-down cDNA during optimization

Figure 8) Plot comparing enrichment difference when using 10 µl of magnetic streptavidin beads or 20 µl

Figure 9) Plots of SL-sequence enrichment and target RNA capture compared to an unenriched mixed cDNA measured by RT-qPCR under a variety of stringency conditions

Figure 10) Gel of completed field sample SL-selected libraries used for analyzing fragment size distribution

Figure 11) TapeStation analysis shows adaptor dimer contamination present in a library prior to clean up

Figure 12) TapeStation analysis of the final pooled library after six consecutive rounds of size selection and clean up with 0.9x mag-bind beads

Chapter 1) Introduction

1.1) General epidemiology and disease burden imposed by African Trypanosomes

African Trypanosomiasis is an infectious vector-borne disease of humans and animals caused by parasitic kinetoplastids of the genus *Trypanosoma*[1]. There are many species of African trypanosome that can infect mammals including *Trypanosoma congolense, Trypanosoma vivax, Trypanosoma simiae,* and the many subspecies of *Trypanosoma brucei*. While all of these are able to infect animals, only *T. brucei rhodesiense* and *T. brucei gambiense* are able to infect humans and cause the disease Human African Trypanosomiasis (HAT) which is also known as sleeping sickness. The WHO estimates that over 60 million people throughout rural areas of sub-Saharan Africa are at risk of developing sleeping sickness. However, government and international interventions to increase sleeping sickness surveillance and treatment have reduced case incidence to less than 10,000 per year[2,3].

African trypanosomes can only be cyclically transmitted by their associated vector, the tsetse fly (*Glossina spp.).* Sustained disease transmission is usually restricted to the vector habitat known as the "tsetse belt" which occupies an area of 8.7 million km^2 throughout tropical and sub-tropical sub-Saharan Africa[2]. The species *T. brucei evansi* and *T. vivax* have adapted to non-cyclical mechanical transmission via the bites of Tabanid and Stomoxyine flies in addition to their usual cyclical tsetse transmission. The emergence of this transmission mode has enabled these two species to spread beyond the tsetse belt of

Africa and establish sustained transmission in other continents. *T. vivax* is established in cattle populations across South America and has the potential to spread further[4].

Animal African Trypanosomiasis (AAT) continues to impose a substantial economic and public health burden on affected areas of Africa to a much higher degree than HAT. The tsetse fly and trypanosomiasis are attributed as a major cause of rural poverty in these areas because they can severely impede or entirely prevent the keeping of cattle and other livestock as all domestic animals can be affected by the disease[2]. In domestic livestock that are introduced into tsetse-infested zones, AAT is a very severe and often fatal disease that can have up to 50% mortality in severe outbreaks[5]. The condition is colloquially known as nagana, which is derived from the Zulu word meaning "powerless" or "useless". Domestic livestock affected by nagana suffer from emaciation, anemia, fever, listlessness, and they become unfit for work as the disease progresses[1]. The combination of the high livestock mortality caused by AAT along with reduced productivity in terms of lowered calving rates, growth rates, milk production, and work output are responsible for an estimated $4.5 billion US in agricultural losses annually[2]. Furthermore, while the main causative agents of disease in cattle are *T. congolense* and *T. vivax*, livestock animals are also able to be infected by human-infectious *T. brucei* and serve as reservoirs for HAT[4].

The drugs currently available for AAT control are considered unsatisfactory due to their toxicity and limited range. Only ethidium bromide, isometamidium, and diminazene aceturate are commonly used and drug resistance is becoming increasingly common, with resistance to one or more of each of these drugs having been reported by 13 sub-Saharan countries[2,5,6]. Ethidium bromide and isometamidium can be used as both curative and

prophylaxis while diminazene aceturate is purely curative in action. In the absence of effective vector control interventions, large-scale prophylactic drug campaigns have allowed for the keeping of livestock in tsetse-infested regions[5]. The control of bovine AAT in particular relies heavily on the use of these drugs. Cattle in affected areas may undergo prophylactic block-treatment periodically at pre-determined intervals, strategic block-treatment once the number of infected individuals in a herd reach a given threshold or are monitored and treated on an individual basis[2]. The implementation of these mass drug administration campaigns comes with some notable difficulties due to lack of transport and frequent drug shortages. There is a pressing need for new prevention and treatment options to combat AAT.

1.2) Antigenic Variation in African Trypanosomes

African trypanosomes live extracellularly in the blood and tissue fluids of their mammalian hosts and are thus constantly exposed to cells of the adaptive and innate immune system, yet they are able to establish chronic infections that can last for several years[7]. The prolonged survival of the parasite within its host is largely dependent on the antigenic variation of a dense variant surface glycoprotein (VSG) coat[7,8,9]. The cell surface of a trypanosome is mainly composed of these GPI-anchored VSGs which are thought to shield invariant surface proteins from immune recognition by either immunodominance or steric hindrance of antibody binding[8]. VSG genes are only expressed from one of 15 telomeric bloodstream form expression sites (BESs) while all others are silent, so only one VSG type may be expressed at a time[8]. The VSG itself is highly immunogenic and induces a strong immune response which contributes to the pathology of trypanosomiasis[2]. At low

antibody titer, VSG turnover through endocytosis at the flagellar pocket prolongs clearance by selectively removing bound antibody from the cell surface[7]. This VSG turnover is insufficient to fully evade the host humoral response and antibodies against the surface glycoprotein coat are effective at killing the parasite at high titers[7]. At this point, *T. brucei* circumvents host antibody recognition by switching the expression of its VSG coat to another antigenically distinct variant[7,8].

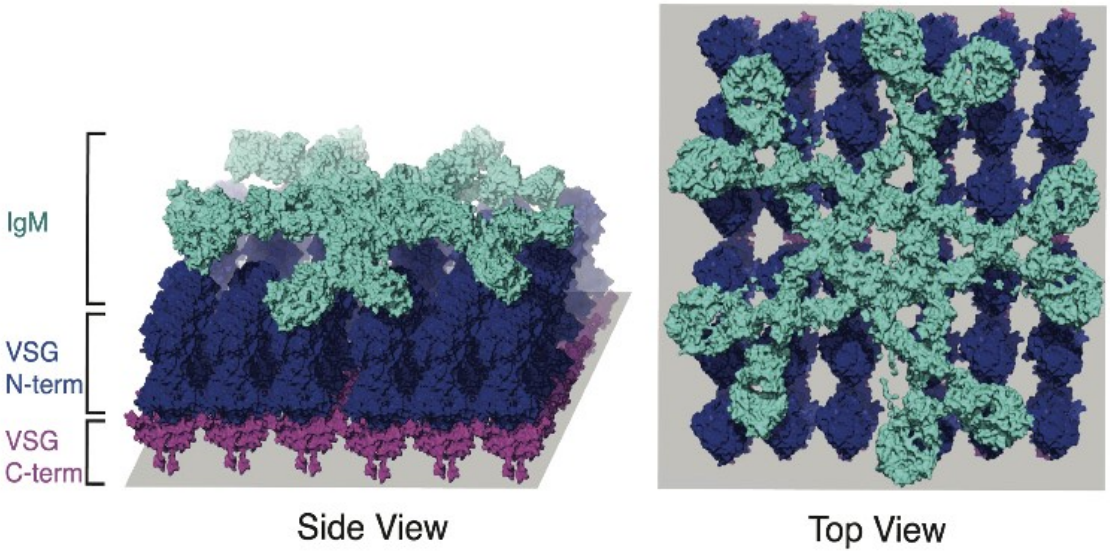


Figure 1) A hypothetical model of the VSG-antibody interaction. The exact arrangement of VSG on the cell surface (grey) and the interaction of VSG dimers with IgM is not known. However, studies have shown that VSG are very densely packed on the cell surface and antibodies are unable to access the C-terminal domains of the VSG (pink)[7,8]. Image from Mugnier et. al. 2016

Switching of VSG expression is thought to occur stochastically, independent of any host factors[7]. Upon induction of switching, the trypanosome population generates cells

expressing many different VSG. Typically, one of these will expand and dominate the circulating population until it is cleared by host antibody and a new variant expands in its place[9]. This dynamic is responsible for the characteristic waves of parasitemia that can be observed in infections[8,9]. There are three ways in which a trypanosome may change the expression of its VSG coat:

1) Transcriptional or *in situ* switching: The active BES is silenced and another BES containing a different VSG gets expressed instead.
2) Gene conversion: A new complete VSG gene replaces the one in the active BES through homologous recombination usually as the result of double stranded DNA breaks and repair. This results in the loss of the previously active VSG and a duplication of the donor.
3) Telomere exchange: Homologous recombination between the ends of two different chromosomes results in a new VSG gene occupying the active BES[7,9].

The evolutionary importance of VSG is made clear by the staggering amount of resources the parasite uses to maintain this coat. VSG makes up 95% of exposed cell surface proteins and about 15% of the total cell protein produced by a single trypanosome with some 1.13×10^7 VSG copies on the surface of each cell at a time[10]. A significant portion of the genome of African trypanosomes is dedicated to keeping up a vast repertoire of different VSG genes for antigenic variation. It has been estimated that the *T. brucei* genome contains ~2700 VSG genes. However, only 20% of these genes encode complete functional proteins[11]. The vast archive of incomplete or pseudogenic VSG genes in the *T.*

brucei genome can be used to generate new VSG antigen diversity through recombination into "mosaic" VSG[12].

1.3) Outstanding Questions

Antigenic variation in African trypanosomes within their mammalian hosts occurs through a very complex system. The antigenic variation of the VSG coat expressed by African trypanosomes is responsible for the parasite's ability to cause chronic infections and is the major reason why there has not been a preventative vaccine developed against trypanosomiasis. Understanding and characterizing the process of VSG switching and expression is key to finding ways to prevent disease. This system is becoming more well characterized in the context of laboratory strains of *T. brucei,* but our current understanding is lacking when it comes to the other animal infectious trypanosome species[3]. Until recently, the lack of laboratory strains of animal infectious trypanosomes such as *T. congolense* and *T. vivax* that could be grown *in vitro* has imposed a barrier to the study of gene expression. Researchers have only just begun to resolve the differences in antigen expression and host/pathogen relationships in these species[13, 14, 15]. Also, the dynamics of VSG expression of any African trypanosome in natural infections is almost entirely uncharacterized. The low parasitemia of natural infections makes isolation of trypanosomes from patient samples not feasible. Transcriptome analyses of African trypanosomes are most often done with laboratory strains grown to a high parasitemia in mice or other laboratory animals[13,16,17]. Parasites are then purified from blood by passage through a cellulose anion exchange column since red blood cells are more negatively charged than parasite cells[18]. This method of parasite purification is not easily adapted to

patient samples or animals with natural infections because the purification does not completely filter out host cells and requires a relatively large amount of blood[13,17].

For my thesis project, I developed a method that allows for the characterization of mRNA expression in wild African trypanosome populations. RNA-seq transcriptome analysis of wild populations could provide insight into the surface antigen expression of parasites in natural infections since mRNA abundance can be used as a surrogate for protein expression. Our approach utilizes the unique spliced-leader sequence present on mature kinetoplastid mRNA and allows us to specifically enrich for parasite mRNA transcripts from whole host blood to use in an RNA-seq library preparation. One group of researchers (Mulindwa et. al.) has performed RNA-seq on trypanosomes from human patient blood and CSF, but their libraries were prepared only after host rRNA degradation. Therefore, this library contained sequences from host mRNA as well as Trypanosome rRNA which were bioinformatically filtered from sequencing reads[19]. Since only a small proportion of the sequencing reads belong to the parasite, these libraries need to be sequenced very deeply in order to resolve any differences in gene expression. The goal of our project was to streamline the library preparation process by depleting host RNA and enriching specifically for parasite mRNA.

Chapter 2) Developing a method for performing RNA-seq transcriptome analysis on field isolates of African trypanosomes

2.1) Introduction and rationale

RNA-seq is a next generation sequencing technique that has become the gold standard for in depth transcriptome analysis. High-throughput sequencing is extremely sensitive, quantitative, and allows the identification of uncharacterized genes when used in combination with transcriptome assembly tools. Typically, the RNA used in the RNA-seq library preparation is extracted from cells or tissues using phenol-chloroform based TRIZol extraction or silica-gel based column procedures[20]. Messenger RNAs are then enriched prior to library preparation either by hybridization with oligo d(T) probes specific for the Poly(A) tails of mature mRNA or selective degradation of ribosomal RNA (rRNA) with exonucleases[21,22]. There is not currently a commercially available rRNA exonuclease degradation kit available for kinetoplastids, so previous trypanosome RNA-seq experiments either use poly(A) hybridization prior to library preparation or they bioinformatically filter rRNA reads after sequencing[19]. This presents a problem for researchers attempting to use RNA-seq to examine the gene expression of parasites from patient samples or field isolates because much of the material in RNA samples from host blood belongs to either the host or rRNA and many of the reads sequenced must be thrown away.

Trypanosomes in natural infections exist in the blood at very low parasitemia, a phenomenon that is apparent considering the widely used diagnostic procedures. Mass

screening for trypanosomiasis in humans is done using a card agglutination test (a serologic test that detects anti-trypanosome antibody) which is very sensitive but lacks specificity[18]. Therefore, other parasitological tests are used to confirm infection. These include techniques such as capillary tube centrifugation, quantitative buffy coat, and mini-anion exchange which all serve to concentrate trypanosomes in blood samples prior to microscopy to confirm diagnosis[18]. Trypanosomes can be grown to relatively high parasitemia in laboratory animals such as mice, rats, or immunosuppressed sheep, but they cannot be reliably isolated from host blood without undergoing multiple rounds of centrifugation and column or chemical treatment[13]. The RNA used to explore gene expression *in vivo* is therefore most often extracted from whole infected host blood and tissue samples. Researchers then make cDNA from their extracted RNA after rRNA depletion by poly(A) hybridization with oligo d(T). Poly(A) tails are an mRNA feature shared by both mammalian hosts and kinetoplastid parasites, so RNA-seq libraries are being prepared with RNA mixtures containing only a small proportion of trypanosome transcripts[17]. The resulting libraries tend to yield a very small proportion of sequencing reads that align to a trypanosome reference while 90% or more belong to the host[19]. Much information can be gained from these libraries, but samples must be sequenced very deeply. Considering the financial cost of high-throughput sequencing, it is wasteful to discard so many of the output reads.

Trypanosomes are early eukaryotes and therefore possess unique molecular features that can distinguish them from their mammalian hosts. They are notably different from other eukaryotes in their mRNA processing mechanisms. Trypanosome genes do not have introns and are arranged in large polygenic clusters that are transcribed in

polycistronic units[26]. The polycistronic pre-mRNA then gets split into individual mature mRNA per gene through post-transcriptional trans-splicing and polyadenylation. During trans-splicing, a conserved 39 nucleotide molecule from the trypanosome spliced-leader RNA (SL-RNA) that carries the 7-methylguanosine cap is donated to the 5' end of every individual mRNA. The spliced-leader RNAs originate from genomic tandem repeats which are transcribed separately from polycistronic units, in contrast to conventional cis-splicing mechanisms[23,24,25]. All of the trypanosomatid pathogens, African trypanosomes, *Trypanosoma cruzi*, and *Leishmania spp*., use this trans-splicing mechanism in mRNA processing. Since the SL-RNA sequence is highly conserved, species specific, and present on the 5' end of all mature expressed mRNA we sought to use the complement of the spliced-leader sequence in place of oligo d(T) in order to jointly deplete host transcripts and rRNA from our RNA samples[26,27].

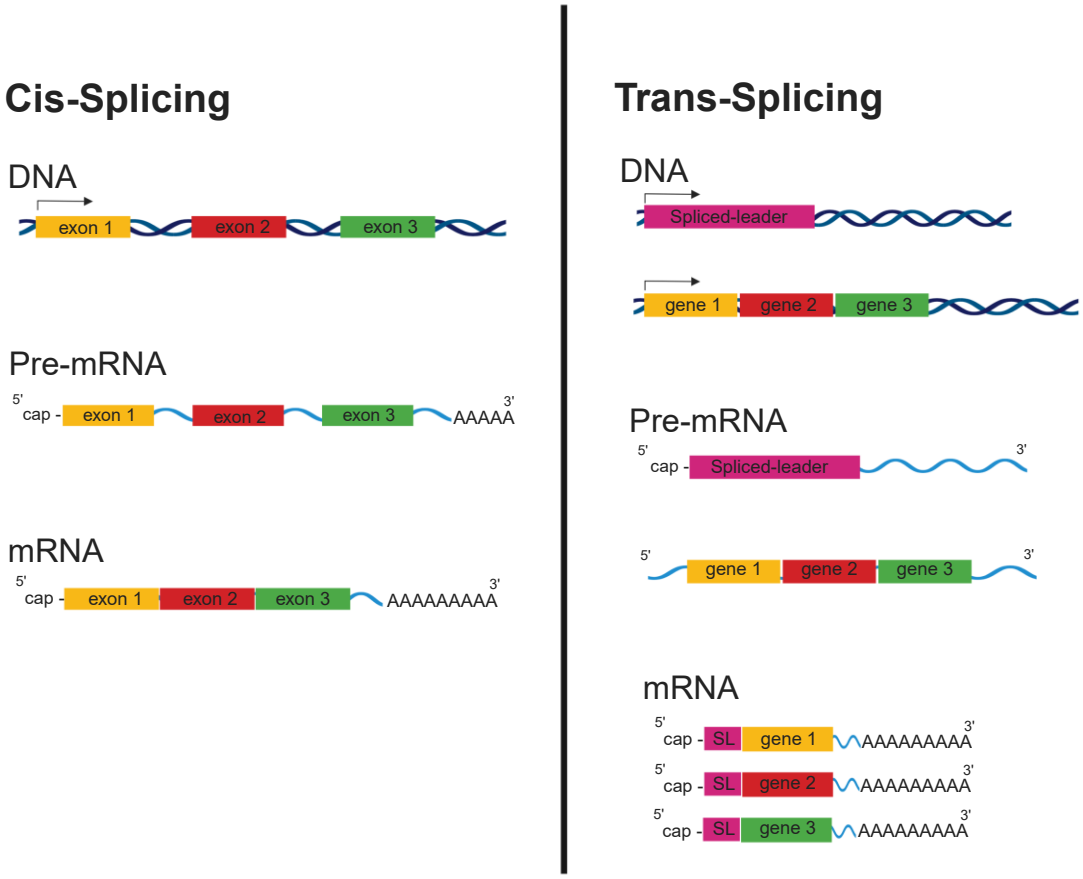


Figure 2) A simplified diagram showing major differences between cis-splicing used during mRNA processing in eukaryotic hosts and the trans-splicing mechanism used by African trypanosomes. Trypanosome genes have no introns and large polycistronic units are dissected into individual mRNA during trans-splicing and polyadenylation. The SL-RNA is transcribed separately and the capped 5' end is donated to the mature mRNA during processing. Image created with Biorender.com.

There is already some precedent for using the SL sequence to study trypanosomes *in vitro.* A proof of concept paper by Gonzalez-Andrade et. al. explores the possibility of using SL-RNA as a molecular diagnostic target for both human and animal cases. They

used reverse transcription of patient RNA samples and real-time PCR with a 19 base-pair primer specific for the trypanosome spliced leader. Their SL-RNA assay was shown to have an analytical sensitivity of 100 parasites/mL of blood and had the advantage of being a surrogate marker for viable organisms, since RNA degrades rapidly after cell death[27]. Another group of researchers has also developed a similar method for SL selection in RNA-seq which they used to sequence *Leishmania donovani* directly from infected tissues without prior purification. The SL trapping protocol used by Cuypers et. al. utilized a combination of random hexamers and SL primer during first strand cDNA synthesis in order to enrich for parasite mRNA for subsequent library preparation. However, the technique has not yet been validated, and it was not compatible with commercially available library preparation kits[26].

2.2) Spliced-leader pulldown

Rather than selectively amplifying trypanosome mRNA during cDNA synthesis using SL-specific primers, we opted instead to replace the oligo d(T) hybridization step of the library preparation with a biotin-streptavidin SL-baited pulldown. The enrichment for parasite mRNA therefore occurs before cDNA synthesis and a commercial Illumina stranded mRNA library preparation kit can still be used. They SL trapping library prep method developed by Cuypers was unable to maintain strand specificity information. We planned to use magnetic streptavidin microbeads and a biotinylated oligo complementary to the SL sequence in place of the oligo d(T) beads that come as part of the usual mRNA selection kit. Binding buffers and salt washes needed to be prepared and optimized for the

streptavidin microbeads. Each species of trypanosome has a slightly different spliced-leader sequence as is shown in table 1.

T. brucei	AACTAACGCTATTATTAGAACAGTTTCTGTACTATATTG
T. congolense	AACTAAAGCTTATAATAGAACAGTTTCTGTACTATATTG
T. simiae	AACTAAAAATTATTATATTACAGTTTCTGTACTATATTG
T. vivax	AACTAAAGCTTTTATTAGAACAGTTTCTGTACTATATTG
T. theileri	AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG
T. cruzi	AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG
T. rangeli	AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG
T. grayi	AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG
Leishmania	AACTAACGCTATATAAGTATCAGTTTCTGTACTTTATTG

Differences to *T. brucei*
Pan-SL-RC-3'biotin oligo

Table 1) The unique spliced-leader sequences for a number of trypanosomatid species. Nucleotide differences between each species and *T. brucei* are highlighted in yellow and the 20-bp common sequence shared by all African trypanosomes is highlighted in blue.

Each spliced-leader sequence as a whole is unique, but the last 20 base pairs are shared for all species of the genus Trypanosoma, including *T. cruzi,* while the *Leishmania Spp.* share a different SL sequence. Our SL pulldown technique could be adapted specifically for mRNA selection of a single species or we could use the common 20 base pair sequence to capture all trypanosome mRNA. This feature is especially helpful in the case of field samples, many of which are often coinfected with multiple species. In order to analyze possible relationships and changes in gene expression due to coinfection with multiple species, we opted to use the 20 bp common sequence for our SL pulldown.

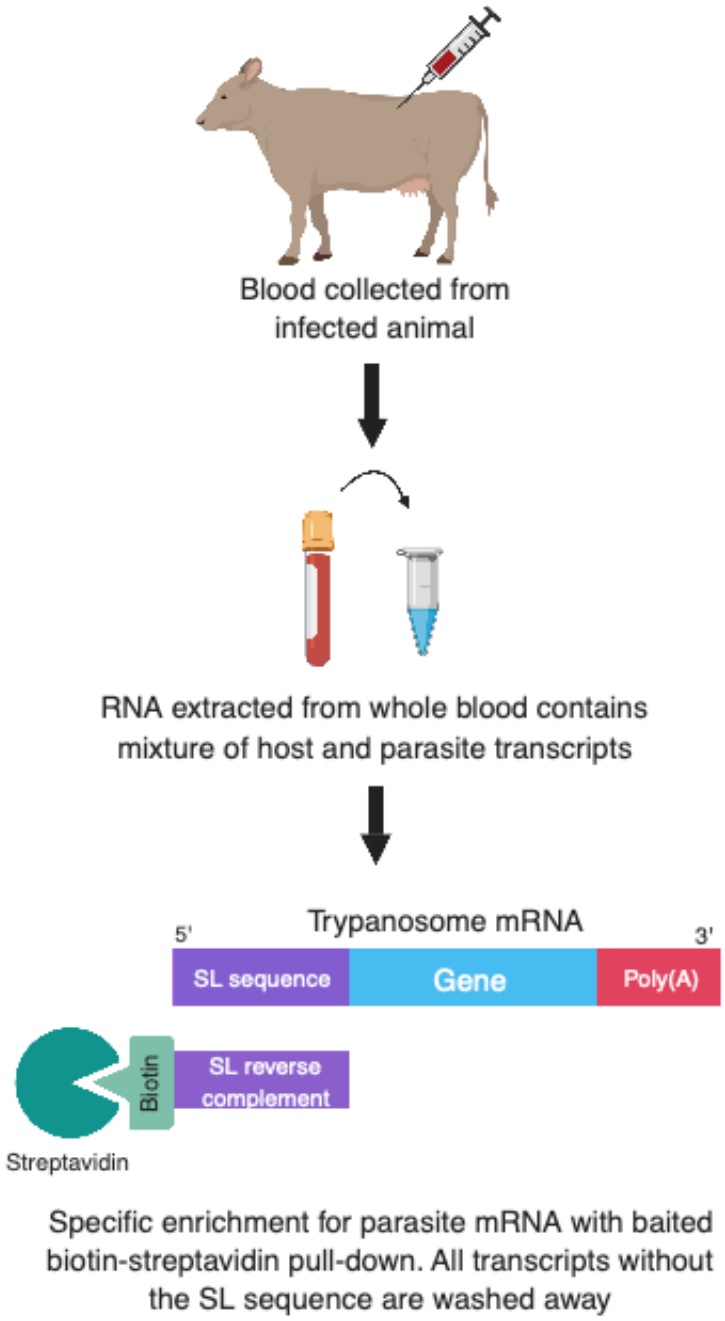


Figure 3) Unlike the poly(A) tail, the spliced-leader sequence is not present on host mRNA. The trypanosome spliced-leader is unique to SL-RNA and processed mature trypanosome mRNA only. By selecting for RNA containing the SL sequence in host/parasite RNA samples collected from natural infections, we can achieve specific

enrichment for parasite mRNA prior to RNA-seq to analyze gene expression in wild populations. Image created with Biorender.com

2.3) Validating the method: comparing oligo d(T) and SL selection

The conventional RNA-seq library preparation isolates poly(A) mRNA using magnetic beads baited with oligo d(T). The isolated mRNA is eluted from the beads, fragmented, and then used to generate double stranded cDNA using random hexamers as primers. Adaptors designed for Illumina sequencing are ligated to the ends of each library cDNA fragment before the whole library is amplified by PCR. The NEB ultra II kit for Illumina contains the reagents needed for all steps following mRNA isolation. The only difference between this method and the SL pulldown is the beads used in the mRNA selection step. I compared sequencing libraries prepared with both methods using known mixtures of mouse and *T. brucei* RNA. Pure mouse and pure *T. brucei* RNA were combined so that there would be either 1% trypanosome or 0.1% trypanosome RNA in a sample containing mainly mouse RNA, to mimic the low parasitemia of RNA extracted from infected blood. 100bp single-end sequencing was performed for all of these experiments and reads were aligned to a *T. brucei* reference genome (version 36) from TriTrypDB using the Bowtie alignment tool[28]. The number of reads aligning to each annotated genomic feature was counted with the HTSeq Python package, and results analyzed in R studio. Under default parameters, Bowtie only reports sequencing reads that uniquely align to one genomic locus. This is important for the accurate quantification of gene expression. The *T. brucei* genome and VSGs in particular are known to contain many repetitive elements which can make it difficult to map 100-bp reads. Therefore, highly similar genes are said

to have low mappability or a low proportion of reads which can uniquely map to them. Gene expression for these experiments was quantified using MULTo, a program specifically designed to correct for the mappability of each gene[29].

The first experiment I performed compared a baseline oligo d(T) library made with 100% trypanosome RNA to two SL libraries prepared with 1% trypanosome, 99% mouse and 0.1% trypanosome, 99.9% mouse RNA mixtures. This experiment tested for enrichment of parasite transcripts when using SL pulldown instead of oligo d(T). The expected output for a library alignment usually reflects the composition of the starting material; however, our results show between a 10 to 20-fold enrichment when using SL selection with more than 10% of reads aligning for the library made from 1% trypanosome RNA and about 1.5% aligning for the one made with 0.1% trypanosome RNA (Figure 4). In contrast, the oligo d(T) library (red bar in Figure 4) was shown to lose some material as only 75% of sequencing reads aligned to the reference even though it was prepared with only trypanosome RNA. This experiment was not suitable for comparing gene expression between SL and poly(A) selection because the RNA composition of the starting material for each library was different, and therefore libraries had variable sizes and amounts of host contamination that would complicate differential expression analysis.

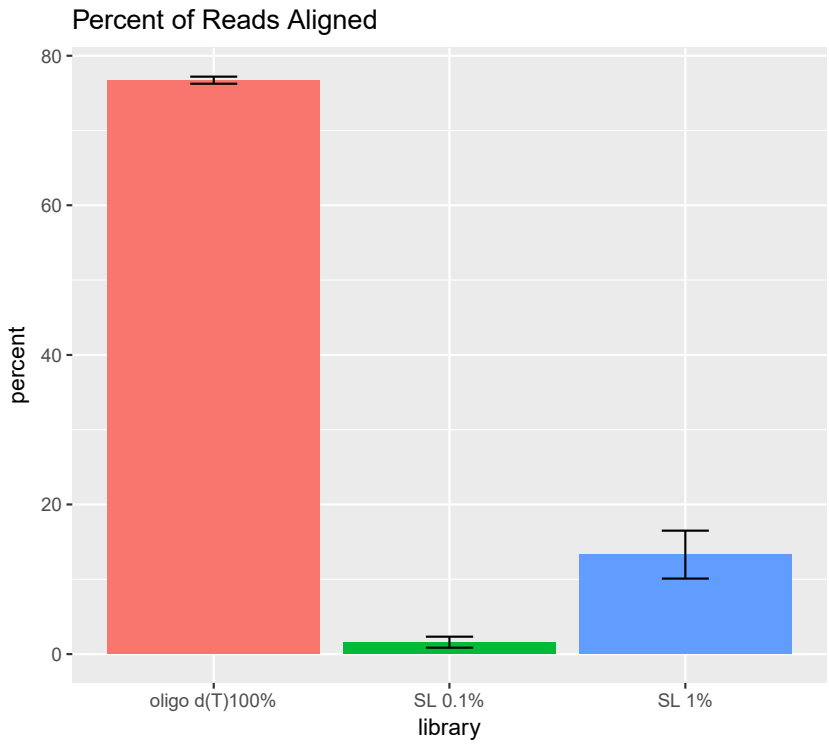


Figure 4) SL sequence-based enrichment of trypanosome mRNA reads. Although this experiment was not suitable for comparing gene expression between methods, it did show enrichment of trypanosome reads. The red bar shows the percentage of sequencing reads from the poly(A)-selected library made with pure *T. brucei* RNA that aligned to the trypanosome reference. The green and blue bars show the percentage of aligned reads for SL-selected libraries made with 0.1% RNA and 1% *T.brucei*/mouse RNA respectively. Percent alignment is expected to reflect the composition of the starting RNA material, so SL-libraries have ~10x the expected number of reads aligning.

To assess if any genes were over or under-represented by the SL library preparation, I also made libraries using both methods with only trypanosome RNA. On the surface, my SL libraries seemed to be outperforming the oligo d(T) method. A higher proportion of

reads were successfully aligning to the trypanosome reference for the SL libraries. However, when I analyzed the gene expression measured by both methods, I found that there was poor correlation between them. I performed two analyses of RPKMs (reads per kilobase of transcript per million mapped reads; a unit of transcript expression that normalizes read counts for genes based on transcript length and library size) calculated using MULTo for the SL and oligo d(T) libraries. One analysis considered only the coding sequences of mRNA, while the other took all genomic features into account. Interestingly, the measured RPKMs for coding sequences when they were analyzed in isolation showed very high correlation (Pearson = 0.97) between gene expression for both types of library while the correlation was poor (Pearson = 0.288) when considering all genomic features. Plotting these RPKM values shows a cluster of genes which are very highly over-represented in the SL libraries. I looked more closely at any gene found to have a log2 fold change greater than 2 to see what kinds of features were responsible for this bias. Not surprisingly, the features with the greatest fold change difference in expression were spliced-leader RNAs. These reads aligning to the spliced-leader sequence did not make up a very large proportion of the total reads aligned, however. They exhibit the greatest fold change difference only because they are typically completely absent from libraries prepared using oligo d(T) selection. Only mature mRNA is polyadenylated and the oligo d(T) selection is unable to capture any noncoding or small RNAs such as the SL-RNA.

We also found that a large number of reads were mapping to rRNA, with an average of 25% of SL library reads aligning to rRNA (Figure 6). However, when all noncoding RNA were excluded from the gene expression analysis, the results of both library preparation methods were highly similar. The discrepancies in gene expression were

almost entirely due to rRNA somehow making it through the mRNA isolation step. I could find no evidence in the literature, or after multiple BLAST searches, that trypanosome rRNA contained the spliced-leader sequence[23,24,30]. Additionally, the three SL library replicates showed variable levels of rRNA as one had a much lower proportion of reads aligning to rRNA than the others. These findings led us to hypothesize that the rRNA detected in the SL library was the result of contamination, or nonspecific binding. In growing mammalian cells, rRNA makes up approximately 80% of the total RNA in a cell and 15% is tRNA, thus mRNA only constitutes a very small portion[31]. Similarly, the vast majority of trypanosome transcripts also belong to rRNAs. From the results shown in figure 6, the largest proportion of SL library reads do in fact align to mRNA coding sequences. The SL selection is depleting rRNA to some degree but not as well as in oligo-d(T) enrichment. To overcome this problem, we needed to increase the stringency of our oligo hybridization conditions in order to decrease off target binding during pulldown enrichment.

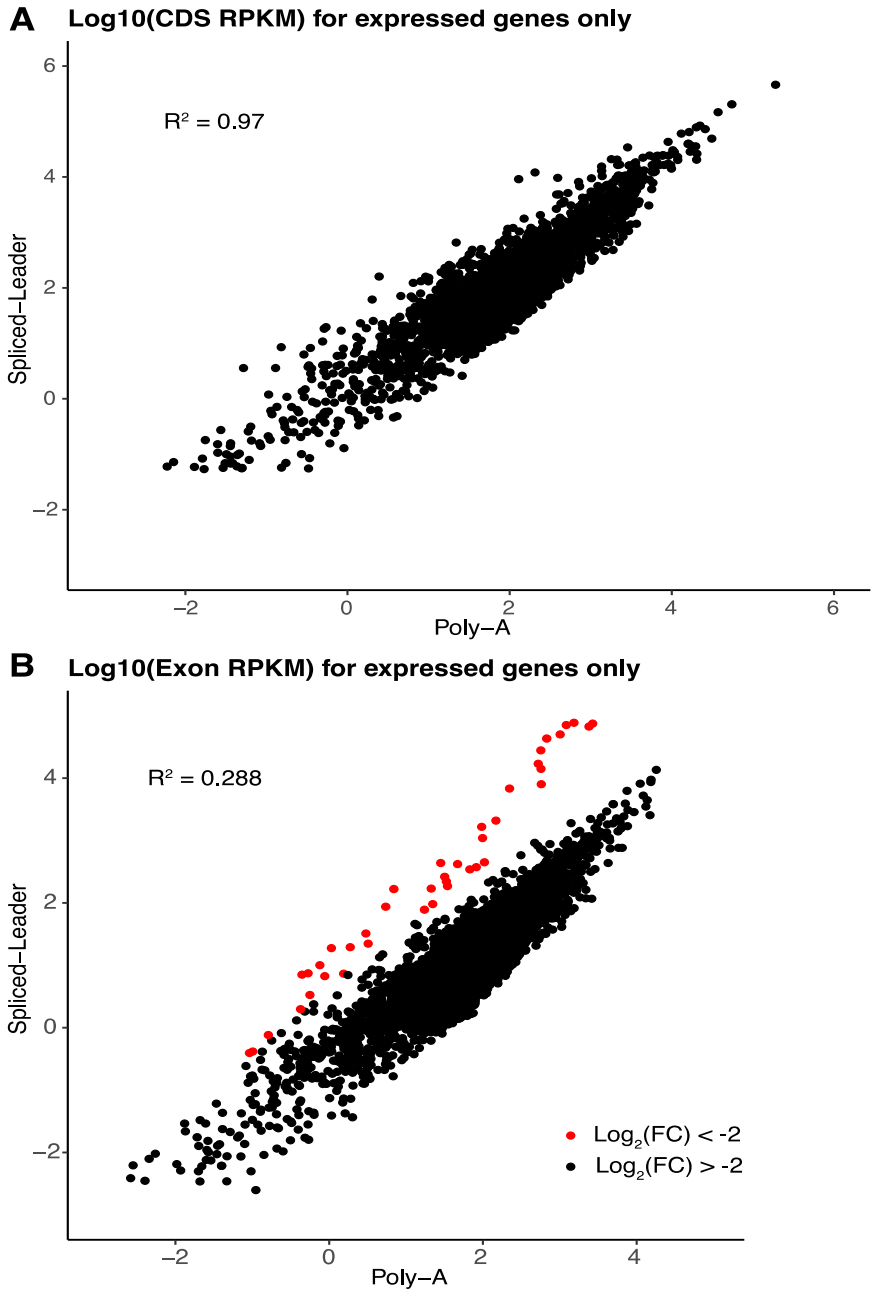


Figure 5) Plots of gene expression RPKMs calculated for oligo d(T) and SL-sequence selected libraries with MULTo for either coding sequences or exons (A) Each point of the plot represents a gene coding sequence. Only genomic features annotated as "CDS" were included in the analysis. The numeric values on the x or y axes correspond to the average RPKM for each coding sequence calculated from spliced-leader or poly-A library

sequencing data. Pearson correlation coefficient is shown. (B) Each point represents an exon, which includes any type of genomic feature such as mRNA, tRNA, rRNA, pseudogene, and noncoding RNA. The numeric values on x and y axes correspond to the average RPKM for each exon. Pearson correlation coefficient is shown and all exons with a greater than 4-fold RPKM bias towards spliced leader are highlighted as red points.

Gene ID	Log_2(FC)	Feature	Product Description
Tb927.10.7650	-2.03546	mRNA	hypothetical protein
Tb927.9.14710	-2.09472	ncRNA	SL RNA
Tb11.v5.0576	-2.09726	mRNA	receptor-type adenylate cyclase
Tb09_rRNA_6	-2.10828	rRNA	M4 ribosomal RNA
Tb10.v4.0129	-2.16356	pseudogene	VSG pseudogene, putative
Tb11.v5.1005	-2.22221	mRNA	hypothetical protein
Tb927.8.6568	-2.23448	tRNA	tRNA glutamine
Tb11.1750	-2.24876	mRNA	hypothetical protein
Tb927.9.14780	-2.34646	ncRNA	SL RNA
Tb11.v5.1011	-2.42666	mRNA	calpain-like cysteine peptidase
Tb11.NT.31	-2.57022	ncRNA	noncoding RNA, putative
Tb927.8.2864	-2.72124	snRNA	small nuclear RNA
Tb927.9.14630	-2.78878	ncRNA	SL RNA
Tb927.4.1213	-2.93516	snRNA	U6 small nuclear RNA
Tb11.15.0008b	-3.04664	mRNA	ESAG3, degenerate
Tb927.9.14870	-3.15846	ncRNA	SL RNA
Tb927.1.5260	-3.37377	mRNA	hypothetical protein
Tb11.v5.1004	-3.42911	mRNA	hypothetical protein
Tb927.11.19810	-3.46773	mRNA	hypothetical protein
Tb927.9.14920	-3.73042	ncRNA	SL RNA
Tb09_rRNA_4	-3.79633	rRNA	M2 ribosomal RNA
Tb927.9.14840	-3.80523	ncRNA	SL RNA
Tb927.8.2861	-3.81974	ncRNA	SRP RNA, 7SL
Tb927.8.449	-3.94945	rRNA	M2 ribosomal RNA
Tb927.9.14810	-3.99335	ncRNA	SL RNA
Tb11.0810	-3.99440	mRNA	hypothetical protein
Tb11.v5.0884	-4.10027	mRNA	hypothetical protein
Tb11.1620	-4.13799	pseudogene	VSG pseudogene, putative
Tb927.2.5680	-4.58549	snRNA	U2 small nuclear RNA
Tb09_rRNA_2	-4.62419	rRNA	M1 ribosomal RNA
Tb927.3.3435	-4.79979	rRNA	M2 ribosomal RNA
Tb927.3.3426	-4.80082	rRNA	M2 ribosomal RNA
Tb927.2.1443	-4.87441	rRNA	5.8S ribosomal RNA
Tb927.2.1540	-4.89220	rRNA	5.8S ribosomal RNA
Tb927.11.19800	-4.94128	mRNA	hypothetical protein
Tb927.11.18720	-5.01779	mRNA	VSG, putative
Tb927.6.185	-5.50345	rRNA	28S beta ribosomal RNA
Tb927.6.187	-5.62014	rRNA	28S alpha ribosomal RNA
Tb09_rRNA_1	-5.63033	rRNA	28S beta ribosomal RNA
Tb927.6.184	-5.65366	rRNA	28S beta ribosomal RNA
Tb927.11.rRNA_2	-5.86331	rRNA	28S alpha ribosomal RNA fragment
Tb09_rRNA_3	-5.99184	rRNA	28S alpha ribosomal RNA

Table 2) The table identifies all of the red points from plot (B) by gene ID. The genes in the table are arranged in decreasing order from least different to most different from poly(A). The features that are most highly overrepresented in spliced-leader are mainly

rRNAs, but there are also some VSG mRNA and pseudogenes that are picked up more by SL-selection.

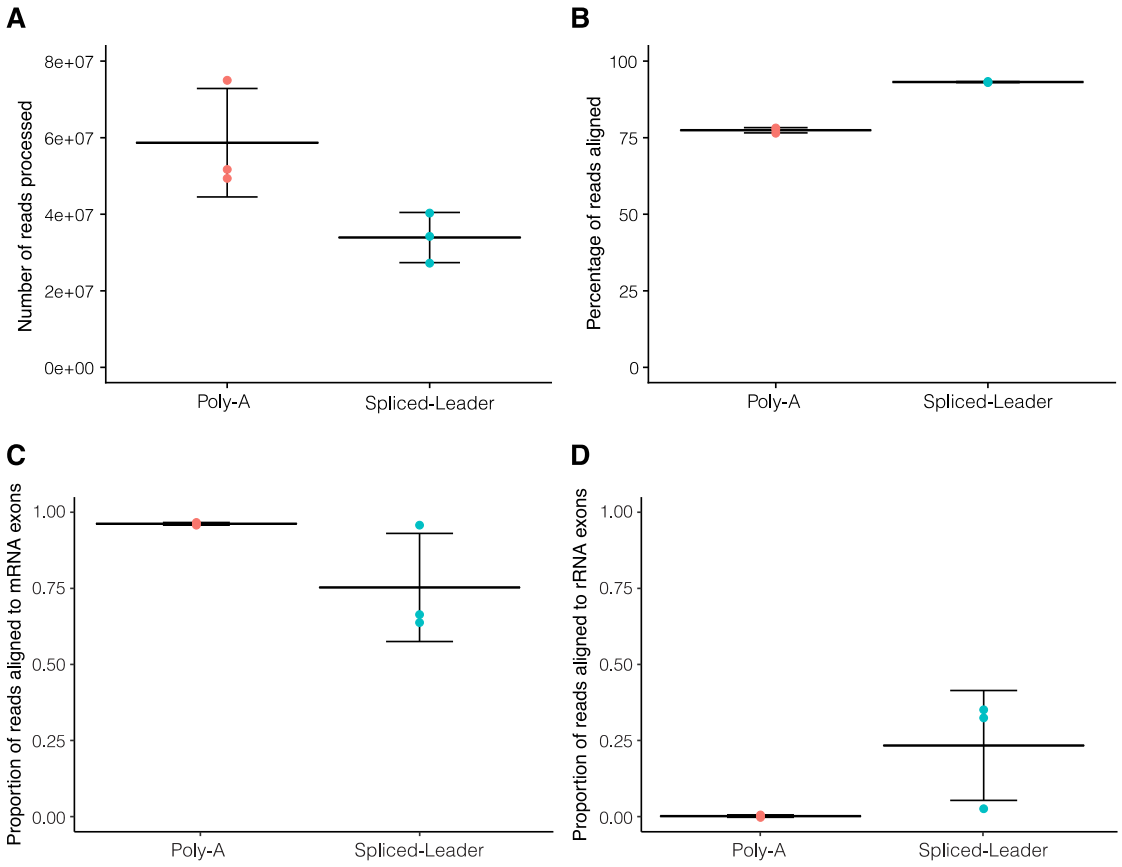


Figure 6) Plots of the total number of reads processed, percentage of reads aligned to the full reference, percentage of reads aligned to mRNA, and percentage of reads aligned to rRNA by library selection type (A) Total number of sequenced reads by library preparation method. (B) The percentage of reads for each method that successfully align to the *T. brucei* TREU927 reference genome as reported by bowtie. This includes both uniquely aligning reads and non-unique ones which may align to multiple loci in the genome. (C) The proportion of successfully aligned reads that align to features annotated as mRNA. (D) The proportion of successfully aligned reads that align to features annotated as rRNA. One of

the SL-selected replicates very closely resembled Poly(A), which led us to hypothesize that contamination was an issue with the current protocol.

2.4) Optimizing SL enrichment

One of my library preparation experiments made with a mixture of 1% trypanosome and 99% mouse RNA included a beads-only control. This is how I found out that the magnetic streptavidin beads alone with no oligo bait still bind enough RNA to create a library that can be sequenced. The library produced by this bead-only sample showed a 1% alignment to trypanosome reference that perfectly reflected the composition of the starting RNA material. In order to improve the enrichment protocol, I needed to prevent nonspecific binding to the magnetic beads and ensure that the baited beads bound strongly to their target sequence. The stringency or specificity of oligo hybridization is dependent upon a number of factors: incubation temperature, incubation time, oligo length, and salt concentration of washes[32,33]. I performed a series of experiments testing the effect of different hybridization stringency conditions on enrichment.

Performing RNA-seq on all of these samples would be too expensive, so we determined the efficiency of the enrichment by RT-qPCR. These experiments quantified relative levels of mouse and trypanosome material present in each sample by amplifying either the trypanosome beta-tubulin or mouse alpha-tubulin housekeeping genes. The abundance of RNA from either host or parasite was estimated using a standard curve generated from serial dilutions of cDNA made with 100% trypanosome or 100% mouse material (Figure 7). I then made experimental cDNA from 1% trypanosome, 99% mouse RNA mixtures that underwent SL pulldown enrichment under a variety of stringency

conditions and compared these to a cDNA control in which no SL pulldown was performed on the RNA before reverse transcription.

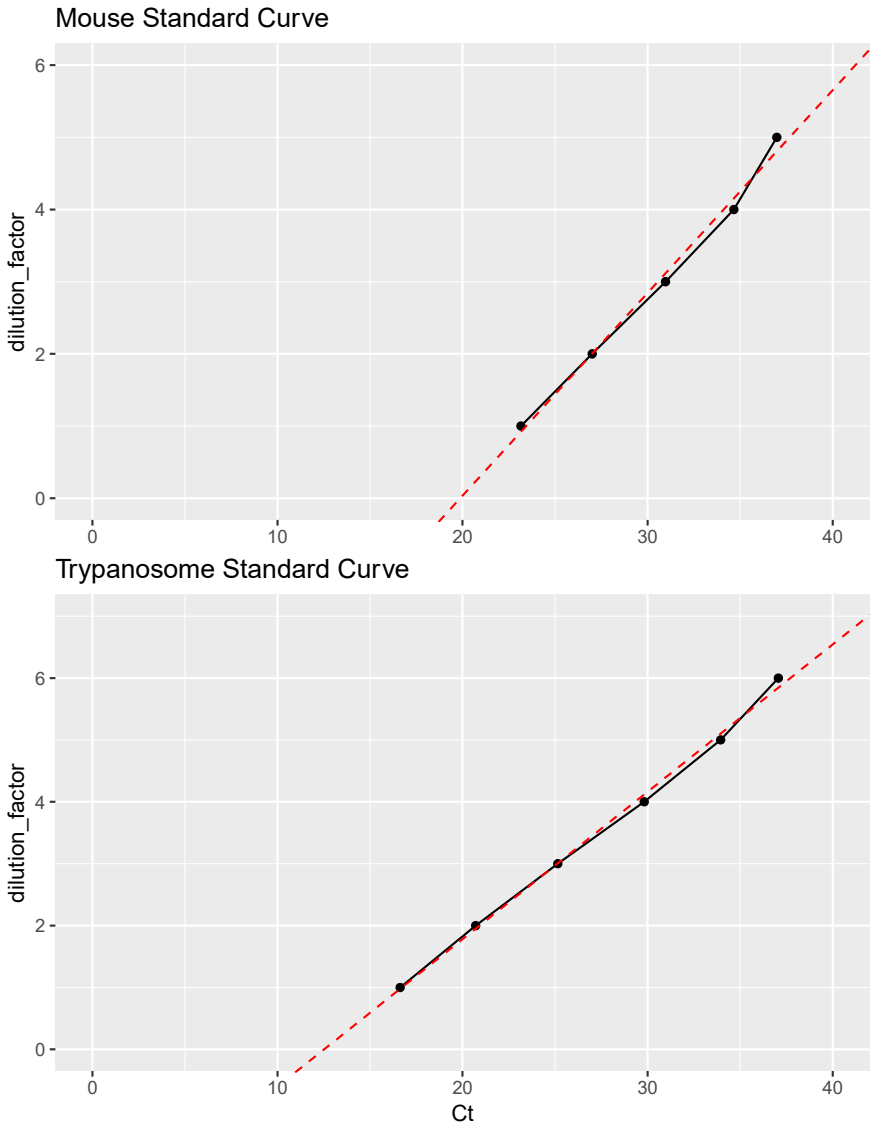


Figure 7) Standard curves generated with RT-qPCR data from serial dilutions of pure *T. brucei* and mouse cDNA. The red dotted line was plotted from a linear regression model calculated from RT-qPCR data points and was used to estimate the abundance of *T. brucei* and mouse material in the SL-enriched 1% trypanosome cDNA samples.

Two changes were made to the standard protocol which we believed would generally improve enrichment by reducing off target binding. Since the nucleic acid binding capacity of the beads very greatly exceeds the amount of target RNA in our mixtures, we decided to reduce the volume of beads used from 20 µl to 10 µl so the samples were not so overloaded with beads. The effect of reducing bead volume is shown in figure 8.

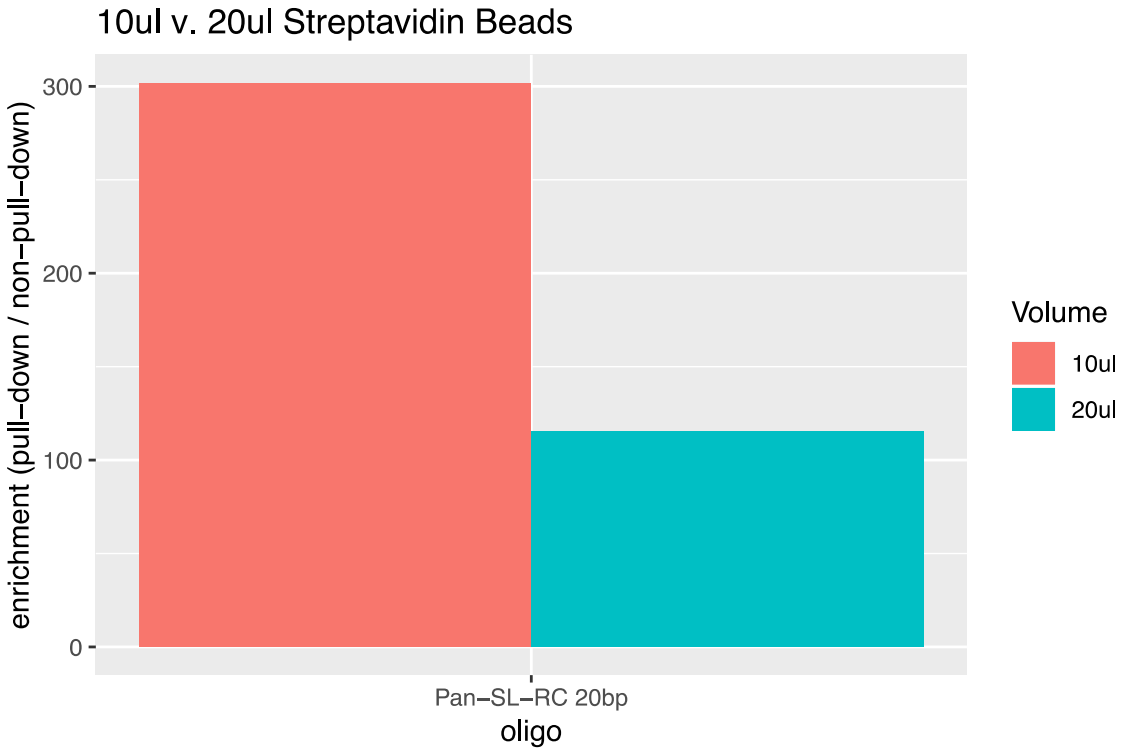
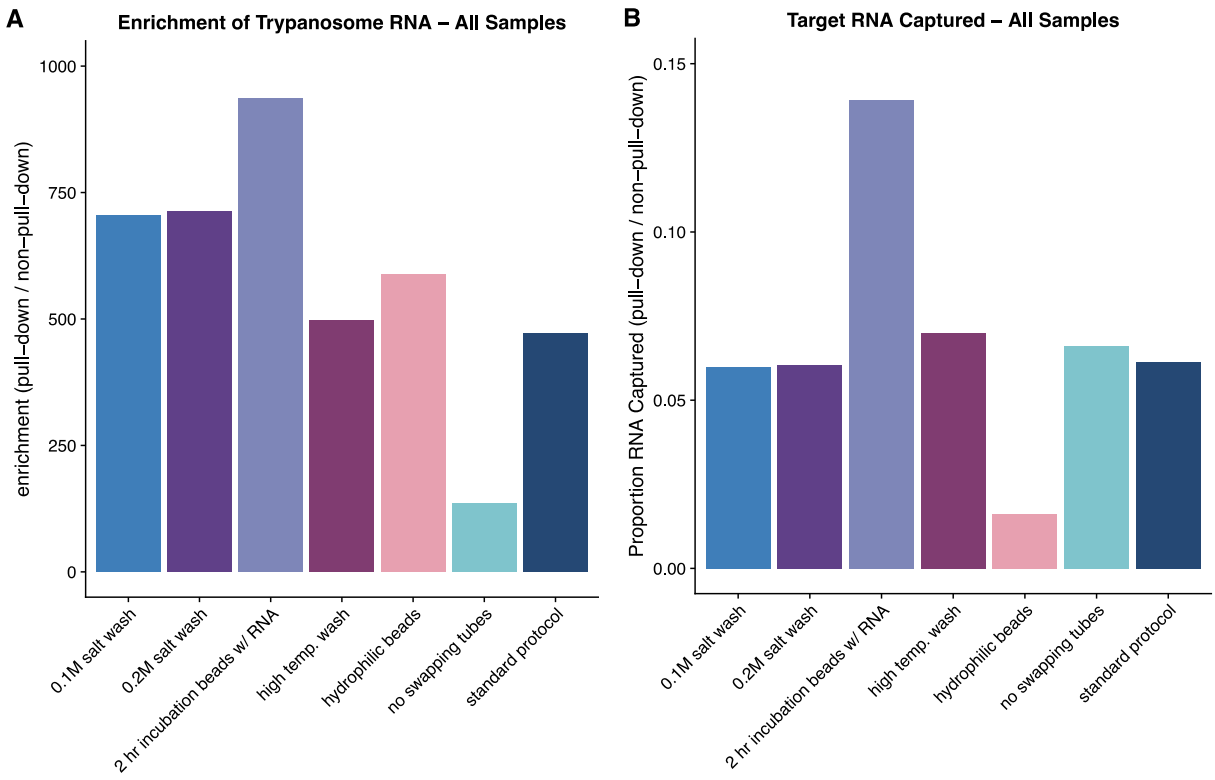


Figure 8) The difference in enrichment found by RT-qPCR when using 10 µl or 20 µl of hydrophilic streptavidin beads in the SL pulldown.

The precise bead volume that would match our target RNA input cannot realistically be used because it would not apply to patient samples where the exact parasitemia is unknown.

To prevent nonspecific nucleic acid binding to the microcentrifuge tube in which the pulldown was performed, we also added a step that moved the beads and pulldown material



into a clean tube before the very last wash.

The standard protocol referred to in figure 9 was as follows:

- 10 µl regular beads (NEB S1420)
- 0.3M NaCl low salt wash
- Ice cold (4°C) low salt wash
- Material moved to clean tubes between washes (swapped tubes)
- 30-minute incubation of oligo-baited beads with RNA

Each condition noted on the x axis indicates how that sample treatment differed from this standard protocol.

Figure 9) Plots of SL-sequence enrichment and target RNA capture compared to an unenriched mixed cDNA measured by RT-qPCR under a variety of stringency conditions. (A) Plot of enrichment comparing the ratio of *T. brucei* to mouse material in experimental pulldown samples under the conditions listed on the x-axis and a no pulldown control. (B) Plot of *T. brucei* RNA captured, calculated as the ratio of *T. brucei* RNA quantified in the experimental enrichment conditions and the no pulldown control.

All of the estimated abundance values for mouse and trypanosome RNA calculated by RT-qPCR with these standard curves (Figure 7) are arbitrary, but their values relative to each other are informative. To compare enrichment, I calculated the ratios of trypanosome to mouse material in the experimental samples as well as the no pulldown mixed cDNA control. The value shown for enrichment represents the ratio of trypanosome to mouse RNA in each sample divided by the ratio measured in a no pulldown control. I also needed to know how much of the starting trypanosome RNA material was captured under each condition. The best set of conditions would simultaneously deplete the most host RNA while also catching the most trypanosome mRNA. I defined target RNA capture as the ratio of trypanosome tubulin mRNA abundance in each condition divided by the abundance value in the no pulldown control.

Incubating the RNA sample with the SL-baited beads for 2 hours resulted in the greatest improvement in both parasite RNA capture and enrichment. The hydrophilic streptavidin beads may have had slightly better enrichment than the regular beads, but the amount of trypanosome RNA captured was much lower. Lower salt concentration in washes makes hybridization more stringent. The 0.1M and 0.2M NaCl salt wash

conditions showed improved enrichment while only slightly affecting capture. The temperature of the wash buffer also makes a difference and high temperature is considered more stringent. Nevertheless, the room temperature wash did not noticeably affect enrichment. One of the most striking differences was when the tubes were not swapped between washes. Not swapping the tubes between washes resulted in the largest decrease in enrichment by far. Although the RNase-free microcentrifuge tubes we use for the preparation reactions are advertised as non-stick, my data shows that quite a lot of the off-target binding was due to RNA material lingering in the tube unbound to the streptavidin beads. Unfortunately, this tube swapping step was added to the standard protocol after the sequenced validation libraries described in chapter 2.3 were prepared. Taking all of these conditions into account, we decided on an optimized SL pulldown protocol:

- 10 µl regular hydrophilic beads (NEB S1420)
- 0.2M NaCl low salt wash
- Ice cold (4°C) low salt wash
- Material moved to clean tubes between washes (swapped tubes)
- 2-hour incubation of oligo-baited beads with RNA

2.5) Application of SL pulldown on RNA from infected cow blood

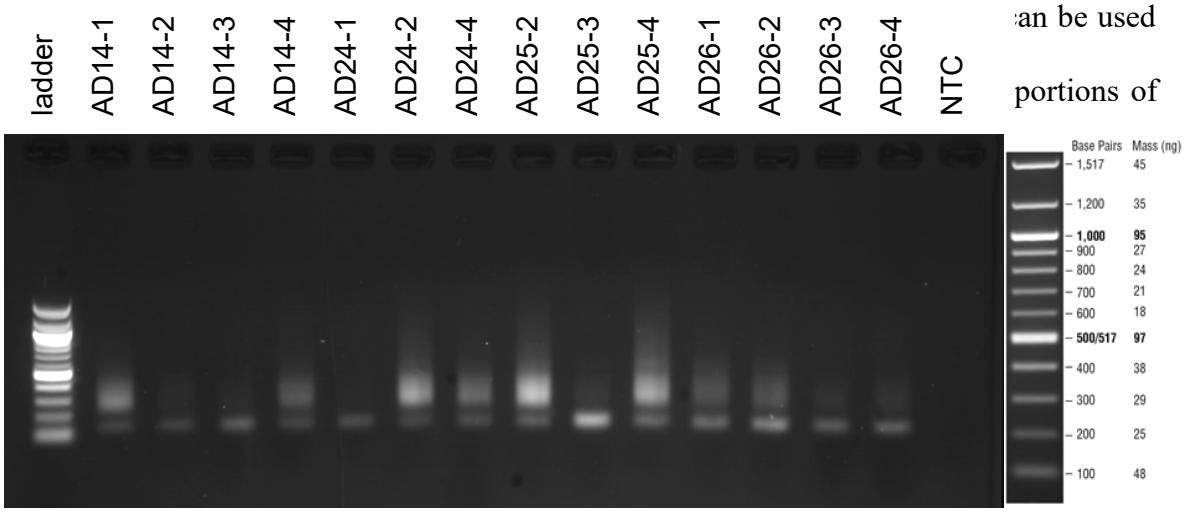
SL libraries prepared under these optimized conditions have not yet been sequenced, but a full library preparation was done on three field isolates of infected cow blood RNA in order to verify that yield under these more stringent conditions is high enough for sequencing. The library yield under our stringent optimized conditions was

between 0.25 and 0.5 ng/µl, much lower than the usual Ultra II yield for 1% trypanosome, 99% mouse RNA mixtures which could be as high as 50 ng/µl. The low yield is likely reflective of the low parasitemia of natural infections, and the more stringent hybridization conditions which should be excluding more off-target RNA. Only 2 nM of library is needed for sequencing, so there was reliably enough to submit.

We have also received over 40 cow blood RNA samples from Ghana, which I have prepared, but we are still waiting for sequencing results. I was concerned about the low yield observed in some of my test libraries, so I made the decision to increase the cycle number of the PCR amplification step of the library preparation protocol from the instructed 15 cycles to 20 cycles. In retrospect, this was probably unnecessary and complicated the library preparations. PCR reactions in general tend to preferentially amplify small fragments, and the reason the NEB kit suggests a maximum of 15 cycles is to prevent the overamplification of contaminating adaptor dimers which are formed when the Illumina adaptor ligates to other adaptors instead of to library fragments. A library that is appropriately prepared and amplified should contain fragments that average 300-bp in size. Adaptor dimers are around 160-bp. The Johns Hopkins GCRF strongly advises that submitted libraries are free of adaptor dimer because the sequencing machine has a severe bias that favors small molecules. Even a small amount of dimer can result in over 60% of the reads being adaptor dimer.

I managed to produce relatively high concentrations of library, with the lowest being 4 ng/µl and the highest 50 ng/µl, but they all contained a significant amount of dimer. I analyzed libraries by gel visualization as well as by TapeStation to determine the composition and size distribution of fragments in each sample. Examples of these results

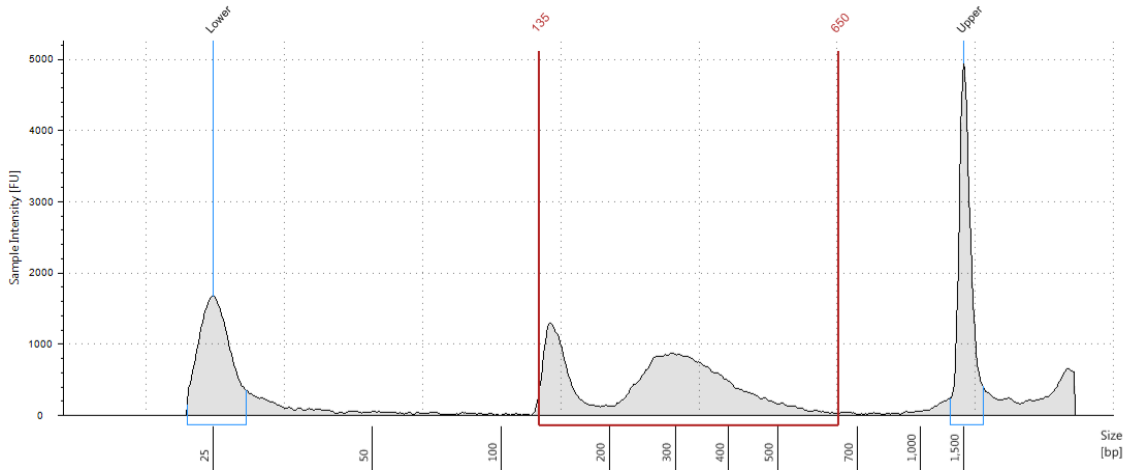
are shown in figures 10-12. High concentration libraries are easier to see by gel. The desired library appears as a smear around the 300-bp ladder mark while the adaptor dimer can be seen clearly as the band below 200-bp. We fixed this issue of dimer contamination though size selection and clean-up with mag-bind beads. Mag-bind beads are magnetic



:an be used portions of every library were pooled based on a regional molarity calculation by the TapeStation analysis and clean-ups were performed on this pool rather than every library individually. The adaptor dimer was sufficiently depleted after six consecutive clean-ups with 0.9x volume of mag-bind beads. This experiment is one of the first attempts to sequence the transcriptome of trypanosomes from natural infections. Our results will be a first look into the gene expression of wild parasites within their natural hosts.

Figure 10) A gel run with 2 µl of completed library immediately following library preparation with the final optimized Ultra II dual index protocol. The expected average library fragment size is 300-bp and shows up on the gel as a smear. The band below this

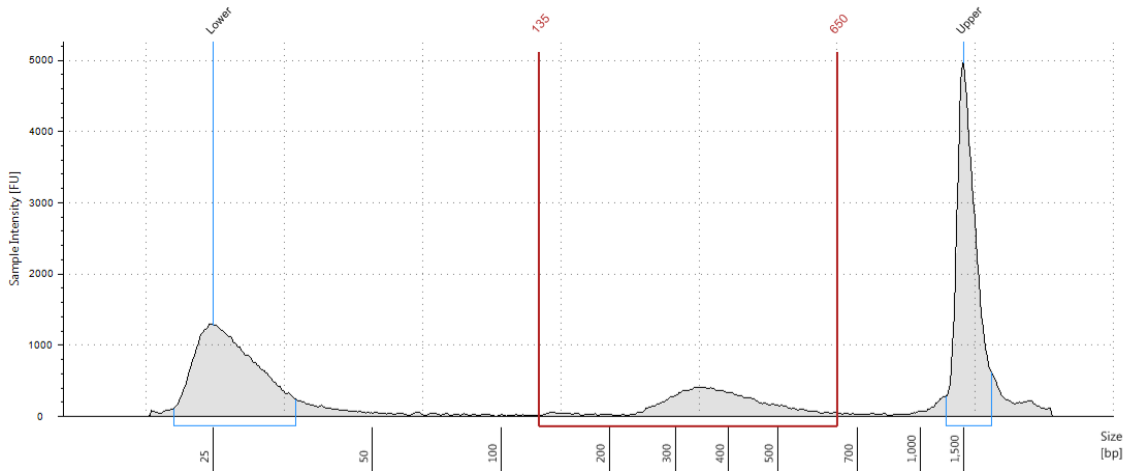
smear is too small to be a library fragment and is indicative of contamination with adaptor dimers. This is a known problem that occurs after too many rounds of PCR amplification, which favors smaller amplicons. High concentration libraries 5 ng/µl or more can be easily visualized by gel, lower concentration libraries can be better analyzed by Agilent TapeStation.



Region Table

From [bp]	To [bp]	Average Size [bp]	Conc. [ng/µl]	Region Molarity [nmol/l]	% of Total	Region Comment
135	650	288	12.5	78.4	71.77	

Figure 11) TapeStation D1000 report and region analysis for a SL-selected library made from RNA extracted from the blood of an infected cow. This analysis illustrates the risks associated with subjecting libraries to too many rounds of PCR, which preferentially amplifies smaller fragments. Library fragments are between 200 to 500-bp in size and the sharp peak around 160-bp is the result of adaptor dimer overamplification.



Region Table

From [bp]	To [bp]	Average Size [bp]	Conc. [ng/µl]	Region Molarity [nmol/l]	% of Total	Region Comment
135	650	378	2.63	11.8	58.24	

Figure 12) TapeStation D1000 report and region analysis of pooled SL-selected libraries made from field isolates after six rounds of size-selection with 0.9x mag-bind microbeads to reduce the abundance of adaptor dimer. A library showing a size profile like this is considered appropriate for sequencing.

Chapter 3) Discussion

These experiments demonstrate the potential of using the unique trypanosome spliced-leader sequence for jointly depleting host RNA and rRNA while enriching for parasite mRNA. In RNA mixtures that contain a minimal amount of target parasite RNA, the SL selective biotin-streptavidin pulldown enrichment increases the proportion of reads in the final library that align to the trypanosome reference by more than 10-fold. Considering that oligo d(T) libraries made from patient samples typically generate libraries with only 1 – 10% alignment to trypanosome reference, this enrichment could prove to be

very useful. Gene expression measured in RNA-seq libraries after SL enrichment is comparable to conventional oligo d(T) selected mRNA libraries. RPKM values calculated for all gene coding sequences shown to have nonzero read counts are highly correlated between the two library types. These results support the validity of SL pulldown as an alternative to conventional poly(A) mRNA selection in RNA-seq library preparations.

While the first SL protocol I described worked well for enrichment and analysis of gene expression, the results of these experiments also showed that there is much room for improvement in the technique. The libraries prepared to examine differential gene expression measurements between SL and poly(A) selected RNAs revealed that there is a problem of off-target capture in the SL preparation. The SL libraries contained many reads that aligned to rRNAs which do not contain the SL sequence and should not have been present after enrichment. Most of the SL sequencing reads did belong to mRNA which demonstrated significant enrichment for our target material, as mRNA make up a minority of total transcripts. Nevertheless, 25% of all reads aligned were from rRNA. We determined that this result was due to nonspecific binding of sample RNA to the magnetic streptavidin beads used in the pulldown. Our beads-only library control revealed that enough RNA could bind nonspecifically to the un-baited beads and tube to produce a full sequencing library. Nonspecific binding to the magnetic beads and tube is likely responsible for the amount of rRNA that made it through the library preparation, and it would also negatively affect enrichment in patient samples through ineffective depletion of host RNA.

In order to resolve the problem of nonspecific binding in the SL pulldown protocol, we performed a number of experiments individually testing the effect of different

stringency conditions on oligo hybridization. Relative amounts of mouse and trypanosome mRNA were quantified by standard curve analysis RT-qPCR using primers specific for mouse or *T. brucei* housekeeping genes. Enrichment efficiency was determined by comparing the ratios of trypanosome to mouse tubulin abundance in pulldown samples versus an unenriched no pulldown control. The proportion of input target RNA captured under each condition was also calculated by comparing the abundance of trypanosome tubulin in the unenriched control and each experimental pulldown sample. The results of this experiment suggested that enrichment of trypanosome mRNA from host/parasite RNA mixtures could be improved by increasing the incubation time of RNA with baited beads and lowering the concentration of NaCl in the final wash. These conditions would ensure that the RNAs binding to the beads were the desired target by requiring stronger binding during oligo-RNA hybridization. Nonspecific binding was also generally improved by reducing the volume of beads used in the pulldown, which made the binding capacity of the beads better match the amount of input target RNA material. Our initial library preparations were overloading the RNA with beads, thus increasing the chances of unwanted nucleic acids binding to them indiscriminately. Additionally, enrichment was greatly improved by moving the whole sample to a clean tube before the final washing step of the pulldown protocol. Much of the off-target RNA material was getting into the library prep by sticking around in the tubes, despite their designation as non-stick RNase-free tubes.

The SL pulldown enrichment protocol was already showing some improvement upon conventional poly(A) mRNA selection before we tried to optimize it. Despite being shown to have problems of contamination, libraries prepared from mixtures of host and

parasite RNA containing minimal amounts of trypanosome material still had more than 10-fold enrichment of parasite sequencing reads. Further experimentation and sequencing of libraries prepared under the optimized conditions is needed to see the degree of improvement in enrichment. Our evidence suggests that spliced-leader selection prior to RNA-seq is a viable alternative to the accepted oligo d(T) method, and it also can increase the amount of sequencing reads that align to the trypanosome reference by excluding host transcripts.

SL selection may be better suited for analysis of certain kinds of gene expression in African trypanosomes, such as VSG expression. Oligo d(T) enrichment is known to come with some important biases, such as higher read density bias towards the 3' ends of genes. Most notably, poly(A) selection is less effective at capturing mRNAs with short tails[22,34]. This probably has a large effect on the RNA-seq transcriptomes of trypanosomes in particular. Since genes are transcribed in polycistronic units, it is thought that regulation of gene expression in trypanosomes mainly occurs post-transcriptionally through differences in the stability and maturation of individual mRNAs[35]. It is therefore highly likely that VSG mRNA, which are transient and must be able to switch their expression quickly, have shorter poly(A) tails and are generally less stable. Poly(A) mRNA selection may very well be less effective at reliably measuring the expression of VSG genes or other short-lived transcripts. My data supports this notion, as some of the genes that were found to be overexpressed in the SL library were VSG mRNA. SL and poly(A) selection target opposite ends of the mRNA, so a combined approach could prove to be useful for understanding transcriptional control and mRNA processing in trypanosomes. SL-mRNA enrichment provides a promising avenue for analyzing gene expression in natural

trypanosome infections, which has not yet been explored due to the limitations of conventional parasite isolation and RNA-seq methods. It can also provide another perspective for looking at transcriptomes of lab adapted parasite strains, since it does not share the same biases as poly(A) selection.

The study of VSG gene expression in wild trypanosome populations is impractical due to the low parasitemia of natural infections, and no in-depth transcriptome analyses have been performed on patient samples. Therefore, whether or not VSG expression behaves the same way in laboratory strains and wild parasites is entirely unknown. By using SL-sequence enrichment in combination with RNA-seq, we can enrich for parasite mRNA transcripts from whole host blood RNA samples. This method allows us to perform RNA-seq with material from a minimal amount of patient blood sample, without first isolating parasites with centrifugation, by depleting host RNA prior to library preparation. Our parasite-enriched transcriptomes could potentially allow researchers to better understand the *in vivo* VSG expression dynamics of parasites in natural infections. It can also easily be adapted to specifically enrich for any species of trypanosome by substituting their unique SL sequences as biotinylated baits. Even African trypanosome species that cannot yet be cultured *in vitro* could be analyzed in this way.

Many researchers have been working to characterize the antigenic variation of the VSG coat expressed by African trypanosomes. Our current understanding of antigenic variation has been resolved mainly by studying *T. brucei,* which has relevance in human disease and has provided a tractable laboratory model for a long time[4]. Until very recently, other species of trypanosome such as *T. congolense* and *T. vivax* could not be cultured in vitro, thus limiting genomic analyses in these parasites[4,13]. Relatively few studies have

focused on the molecular biology of *T. vivax* and *T. congolense*, but those that have been done show many compelling differences in VSG expression and genomic structure between the different species. *T. vivax* in particular is the most ancestral of the African trypanosomes and has VSG genes with significantly different properties than its relatives: two unique protein subfamilies not inherited by *T. brucei* and *T. congolense*, and a reduced rate of recombination[14,15]. Since recombination is responsible both for VSG expression switching and mosaic VSG formation, it is possible that the *T. vivax* VSG repertoire has a reduced capacity for generating antigenic diversity and the surface antigen expression of this species may be relatively stable. Transcriptome analysis of a laboratory strain of *T. vivax* has also suggested that VSG represent only 55% of the surface proteome of *T. vivax,* versus 95% for *T. brucei*[13]. This finding is further supported by electron microscopy observations of the cell surface of *T. vivax*, which qualitatively appears to have less dense coats than *T. brucei*[36]. It is possible that the VSG coat of *T. vivax* cannot efficiently act as a protective barrier against host immunity as it does in *T. brucei* and may indicate functional differences between the VSG coats of the trypanosome species in general. Moving forward, our lab is interested in using SL-enrichment and RNA-seq to characterize the mRNA expression of *T. vivax* and make inferences regarding the surface protein expression of this species. By analyzing gene expression in wild parasites, we may be able to resolve how the parasite is able to maintain chronic infections and gain insight into the evolution of antigenic variation.

Chapter 4) Methods

4.1) *T. brucei* cell culture and RNA extraction

Pure trypanosome RNA was used for control libraries and to make known 1% and 0.1% mixtures of trypanosome and mouse RNA. RNA was extracted from *T. brucei* Lister 427 single marker cells grown *in vitro* in HMI-9 media at 37°C[37]. The entire volume of a parasite culture containing 50 million cells was centrifuged for 10-minutes at 1500 rpm to pellet the parasite cells. The media supernatant was discarded by pouring, parasites were resuspended in any remaining media and moved to a 1.5 mL microcentrifuge tube for further isolation. Cells were spun again at 5200 rpm for 4 minutes and the remaining HMI-9 supernatant removed. Parasite cells were immediately resuspended in 1 mL of TRIZol for RNA extraction according to manufacturer's instructions.

All traces of genomic DNA must be removed from the RNA sample prior to library preparation. Pure *T. brucei* RNA control samples were treated immediately following extraction from culture using TURBO DNase according to manufacturer's instructions, besides the inactivation step. Instead of inactivation with the reagent provided in the kit, a clean-up with 1.8x magnetic microbeads was performed in order to prevent contamination with any residual reagents. After DNase treatment, each sample was checked for DNA contaminants by PCR using OneTaq polymerase (NEB M0482) and *T. brucei* beta-tubulin primers.

Tryp Tubulin-F: GAACCACTTGGTGTCTGCTG

Tryp Tubulin-R: TAGCTCGGGCACGGAGAGA

4.2) Mouse RNA controls

The mouse RNA used in the known mixture controls for qPCR and library preparations were extracted from uninfected homogenized mouse brain tissue using TRIZol according to manufacturer's instructions. Pure mouse and RNA control samples were treated immediately following extraction from tissue using TURBO DNase according to manufacturer's instructions, besides the inactivation stepwhich was replaced with a clean-up with 1.8x magnetic microbeads in order to prevent contamination. After DNase treatment, each sample was checked for DNA contaminants by PCR using OneTaq polymerase (NEB M0482) and mouse genomic tubulin gene primers.

MsTub-F: ATCTCCATCCATGTTGGCCA

MsTub-R: GGTCAATGATCTCCTTGCC

4.3) Mag-bind bead clean-ups

Many steps in the library preparation protocol require purification of RNA and DNA material. The volume of beads used relative to the total sample volume can also be used for fragment size selection purposes as well as nucleic acid purification. All magnetic bead clean-up steps used Mag-Bind TotalPure NGS (Omega Bio-tek: M1378-01). 1.8x bead volume captures the entire sample and was used for the clean-up of RNA samples following DNase treatment while 0.9x bead volume will preferentially bind larger molecules and leave small > 200-bp fragments behind in the supernatant.

Clean-ups with the mag-bind beads generally followed this procedure. The appropriate size-selective volume of beads (1.8x for all, 0.9x for large only) was added to the RNA or library cDNA sample. Beads and nucleic acid sample were mixed by gently

pipetting up and down about ten times and then incubated at room temperature for 15 minutes without shaking. The tube was then placed on a magnet for 2 minutes so the beads could pellet on the side of the tube. The supernatant is removed and discarded. Without removing the tube from the magnet, the bead pellet was washed twice with 200 µl of 80% ethanol solution which must be made fresh prior to each clean-up. The clean material was eluted from the beads by removing the tube from the magnet, adding nuclease-free water or 0.1X TE buffer, gently pipetting up and down several times, and then replacing the tube on the magnet for 2 minutes to pellet beads. The eluted volume containing the desired material was transferred to a clean tube.

4.4) SL and oligo d(T) library preparations for validation

Testing Enrichment:

The libraries showing enrichment of parasite sequencing reads in SL-selection compared to oligo d(T) were prepared using the NEBNext Ultra Directional Library Preparation kit for Illumina (E7420) which can take a maximum input of 2 µg of RNA starting material per library preparation reaction. The kit was used in combination with NEBNext multiplex oligos for Illumina index primer set 1 (E7335) which provides unique indexes to library fragments for single-end sequencing. 6 µg known mixtures of mouse and trypanosome RNA were prepared in order to make 3 technical library replicates for each, one containing 1% and another with 0.1% trypanosome RNA. Poly(A) mRNA isolation was performed on 2 µg of pure trypanosome RNA using NEBNext Oligo d(T)25 beads and the accompanying kit wash buffers according to manufacturer instructions. Washing and

binding buffers for the SL-enrichment were prepared as well as stock solution of biotinylated SL oligo (sequence highlighted in table 1)

SL Pulldown Buffers:

2x Wash/Binding Buffer [1.2 M NaCl, 40 mM Tris-HCl(pH 7.5)]:

- 1 ml (25 µl per prep)– 240µl 5M NaCl, 40 µl 1 M Tris-HCl (pH 7.5), 720 µl water

Low Salt Buffer [0.3 M NaCl, 20 mM Tris-HCl (pH 7.5)] :

- 1 ml (100 µl per prep) - 60 5M NaCl, 20 µl 1 M Tris-HCl (pH 7.5), 920 µl water
- keep on ice

1x Wash/Binding Buffer [0.6 M NaCl, 20 mM Tris-HCl(pH 7.5)] :

- 10ml (525 µl per prep) - 1.2ml 5M NaCl, 0.2 ml 1 M Tris-HCl (pH 7.5), 8.6 ml water

Pan-SL-RC-biotin oligo: stock at 1 nmol/µl in 10mM Tris-HCl (pH 7.5)

Oligo solution master mix was made by mixing 24 µl of 1x wash/binding buffer with 1µl of 1 nmol/ul biotinylated SL oligo stock per reaction. SL mRNA isolation was done using 20 µl of NEB hydrophilic magnetic streptavidin beads (S1421) for each reaction which were washed with 100 µl of 1x wash/binding buffer and resuspended in 25 µl of oligo solution. The streptavidin beads and biotin-tagged oligo are incubated together for 5 minutes gently shaking. They are then applied to a magnet and washed twice with 100 µl 1x wash/binding buffer. During this time, the RNA sample was prepared. The 2 µg of RNA were brought up to a volume of 25 µl with water and then 25 µl of 2x wash/binding buffer

added. Secondary RNA structure is dissolved by incubating the RNA sample at 65 ºC for 5 min and then placed on ice for 3 minutes to cool before proceeding with reaction. The total RNA sample was added to the prepared magnetic beads and incubated at room temperature for 30 minutes gently shaking. After incubation, each sample is then washed twice with 100 µl of 1x wash/binding buffer and once with 100 µl of ice cold 0.3M low salt buffer. All of the supernatant must be removed prior to elution and fragmentation so each tube was quickly spun down and residual wash buffer removed with a 10 µl pipette tip. SL-selected RNA was eluted from the beads by adding 15.5 µl of first strand synthesis reaction buffer and priming mixture from the NEBNext Ultra Directional Library Preparation kit for Illumina (E7420), incubating at 94°C for 15 minutes, applying to a magnet, and collecting 10 µl of eluate. At this point, the remainder of the library preparation is done according to manufacturer instructions and does not differ from the oligo d(T) library preparation.

Final libraries were analyzed for appropriate fragment size and absence of adaptor dimer using the Agilent 2200 TapeStation high sensitivity D1000 ScreenTape system. The expected average fragment size for libraries prepared with the NEB Ultra kit is about 300-bp. Libraries free of dimers were quantified by RT-qPCR CT standard curve using standards and reagents from the KAPA Library Quantification Kit. Libraries submitted for sequencing must be pooled together such that each library represented in the pool has the same number of molecules present. Once the molarity of each library was calculated from the KAPA RT-qPCR standard curve, all samples were diluted to match the lowest concentration sample. Equimolar portions were pooled together and then diluted to 2nM for submission to the Johns Hopkins Genomics Core Research Facility (GCRF) for

sequencing. The Libraries were sequenced on an Illumina HiSeq 2500 producing 100-bp single-end reads.

Testing Gene Expression:

The SL and oligo d(T) libraries made with pure trypanosome RNA were prepared with the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760) in combination with NEBNext multiplex oligos for Illumina index primer set 1 (E7335). The Ultra II kit was an improved version of the NEB Ultra Directional kit that was shown to have better yields. The Ultra II kit was used for RNA library preparations from this point onwards. The recommended maximum RNA input for the Ultra II kit was 1 µg. Only pure *T. brucei* RNA was used as input for these library preparations. Three technical replicates were prepared using oligo d(T) or SL selection. The oligo d(T) and SL selection protocols did not differ from the enrichment experiment and the library preparation was done according to manufacturer instructions.

The NEB Ultra II kit produced very high library yields. Average library fragment size could be assessed visually by gel electrophoresis running 2 µl of the library instead of using the high sensitivity TapeStation. Libraries free of adaptor dimers were quantified by RT-qPCR CT standard curve using standards and reagents from the KAPA Library Quantification Kit. Libraries were then pooled, diluted to 2nM, and submitted to the Johns Hopkins GCRF for 100-bp single-end sequencing on the Illumina HiSeq2500.

4.5) Enrichment optimization

Stringency of oligo hybridization is affected by temperature, salt concentration, oligo length, and incubation time. High temperature and low salt concentration are considered more stringent[32,33]. A qPCR assay was developed to test the effect of more stringent conditions on trypanosome RNA enrichment. The SL pulldown was performed on 1 µg of a 1% mixture of *T. brucei* and mouse RNA as if it was going to be used in a library preparation. However, the RNA was eluted from the streptavidin beads with warm 65 °C water instead of with the first strand synthesis mixture. 8 µl of RNA eluate was then used to make single-stranded cDNA primed with random hexamers using Superscript III kit (Invitrogen 18080 - 051).

The first change analyzed was reduced volume of beads used in the pulldown. The same conditions as the original library preparations were used and the only variable condition was the volume of beads, which was reduced from 20 µl to 10 µl of hydrophilic beads. Results are shown in figure 8. The remainder of optimization experiments were performed using 10 µl of beads because of the difference found in this experiment.

SL Pulldown Buffers:

Common Buffers:

<u>2x Wash/Binding Buffer [1.2M NaCl, 40mM tris-HCl (pH 7.5)]:</u>

- 1 ml (25ul per prep) – 240ul 5M NaCl, 40ul 1M Tris-HCl (pH 7.5), 720ul water

<u>1x Wash/Binding Buffer [0.6M NaCl, 20mM Tris-HCl (pH 7.5)]:</u>

- 10ml (525ul per prep) – 1.2ml 5M NaCl, 0.2 ml1M Tris-HCl (pH 7.5), 8.6ml water

Oligos:

Pan-SL-RC-biotin oligo: stock at 1 nmol/µl in 10mM Tris-HCl (pH 7.5)

Final Wash Buffers:

Standard protocol wash [0.3M NaCl, 20mM Tris-HCl (pH 7.5)]:

- 1ml (100ul per prep) – 60ul 5M NaCl, 20ul 1M Tris-HCl (pH7.5), 920ul water

Medium salt wash [0.2M NaCl, 20mM Tris-HCl (pH7.5)]:

- 1ml (100ul per prep) – 40ul 5M NaCl, 20ul 1M Tris-HCl (pH7.5), 940ul water

Low salt wash [0.1M NaCl, 20mM Tris-HCl (pH7.5)]:

- 1ml (100ul per prep) – 20ul 5M NaCl, 20ul 1M Tris-HCl (pH7.5), 960ul water

Sample	Oligo	Incubation	Low NaCl Washes	Final Wash Temp.	Beads
No pulldown	none	n/a	n/a	n/a	none
Standard Protocol	Pan-SL-RC	30 min	0.3 M	4°C	Regular
No tube swapping	Pan-SL-RC	30 min	0.3 M	4°C	Regular
Medium salt wash	Pan-SL-RC	30 min	0.2 M	4°C	Regular
Low salt wash	Pan-SL-RC	30 min	0.1 M	4°C	Regular
High temp. wash	Pan-SL-RC	30 min	0.3 M	20°C	Regular
Long incubation	Pan-SL-RC	2 hr	0.3 M	4°C	Regular
Hydrophilic beads	Pan-SL-RC	30 min	0.3 M	4°C	Hydrophilic

Table 3) A detailed list of hybridization conditions. Sample name indicates which set was the control, which was considered the standard set of conditions, and how each sample differs from the standard specifically.

Each sample was treated differently according to specifications outlined in table 3 above, but they followed the same overall protocol. Oligo solution master mix was made by mixing 24 µl of 1x wash/binding buffer with 1µl of 1 nmol/ul stock per reaction. SL mRNA isolation was done using 10 µl of either regular NEB magnetic streptavidin beads (S1420) or NEB hydrophilic streptavidin beads (S1421). The beads aliquoted for each reaction were washed with 100 µl of 1x wash/binding buffer and resuspended in 25 µl of oligo solution. The streptavidin beads and biotin-tagged oligo are incubated together for 5

minutes gently shaking. They are then applied to a magnet and washed twice with 100 µl 1x wash/binding buffer. During this time, the RNA sample was prepared. The 1 µg of RNA was brought up to a volume of 25 µl with water and then 25 µl of 2x wash/binding buffer added. Secondary RNA structure is dissolved by incubating the RNA sample at 65 ºC for 5 min and then placed on ice for 3 minutes to cool before proceeding with reaction. The total RNA sample was added to the prepared magnetic beads and incubated at room temperature for the specified incubation time while gently shaking. Each sample is then washed twice with 100 µl of 1x wash/binding buffer, moved to a clean microcentrifuge tube, and washed once more with 100 µl of low salt buffer. To elute SL-selected mRNA from the beads 15 µl of warm 65°C water was added to the beads, incubated for 2 minutes at 65°C in a heat block, applied to a magnet, and 8 µl of eluate collected for use in cDNA synthesis reaction. The no pulldown control was made from 8 µl of the 1% *T. brucei* and mouse RNA mixture primed with random hexamers using the Superscript III kit (Invitrogen 18080 - 051). Pure *T. brucei* and mouse cDNA were also made using the superscript III kit in order to make standard curves to estimate relative levels of trypanosome and mouse RNA in the experimental cDNA samples.

The composition of single stranded cDNA was determined using qPCR with Applied Biosystems SYBR Green PCR Master Mix (4309155). Mouse and trypanosome material were detected in each sample using primer probes specific for either mouse alpha tubulin or trypanosome beta-tubulin housekeeping genes[38].

Mouse TubA-F: TGTCCTGGACAGGATTCGC

Mouse TubA-R: CTCCATCAGCAGGGAGGTG

T. brucei Tubulin-F: GAACCACTTGGTGTCTGCTG

T. brucei Tubulin-R: TAGCTCGGGCACGGAGAGA

The 20 µl reactions were prepared and each sample run in duplicate. 1 µl of cDNA template combined with 10 µl SYBR Green Master Mix, 0.1 µl of 10 µM forward primer, 0.1 µl of 10 µM reverse primer, and 8.8 µl water. Serial dilutions of pure trypanosome and pure mouse cDNA from 1:10 to 1:1,000,000 were made and used to generate standard curves. By using a linear regression model with CT values found for each dilution in the standard curve, the relative abundance of trypanosome and mouse material in each experimental sample could be estimated from their measured CT values.

4.6) SL-sequence enrichment and RNA-seq

The NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760) was used. To keep the samples within the working range of this kit, no more than 1 µg of target RNA should be used. However, RNA samples isolated from whole blood likely contain a very small fraction of trypanosome RNA and the exact parasitemia for each sample unknown. 10 µl of patient RNA sample was used as input for each library preparation. None of these contained more than 5 µg of RNA so we assumed that the amount of Trypanosome RNA present in each sample was within the working range of the library kit.

Paired-end sequencing gives more options in terms of sequencing depth and can be more cost effective. We planned to start performing paired-end sequencing on SL libraries using different Illumina machines that could produce more reads. The many models of Illumina sequencers have different requirements regarding the single and dual

indexes used to distinguish different libraries during sequencing. For these preparations, the Ultra II kit was used in combination with NEBNext multiplex oligos for Illumina dual index primer set 1 (E7600), so that libraries could be compatible with the Illumina NovaSeq.

SL Pulldown Buffers:

2x Wash/Binding Buffer [1.2 M NaCl, 40 mM Tris-HCl(pH 7.5)]:

- ml (25 µl per prep)– 240µl 5M NaCl, 40 µl 1 M Tris-HCl (pH 7.5), 720 µl water

Low Salt Buffer [0.2 M NaCl, 20 mM Tris-HCl (pH 7.5)] :

- 1ml (100ul per prep) – 40ul 5M NaCl, 20ul 1M Tris-HCl (pH7.5), 940ul water
- keep on ice

1x Wash/Binding Buffer [0.6 M NaCl, 20 mM Tris-HCl(pH 7.5)] :

- 10ml (525 µl per prep) - 1.2ml 5M NaCl, 0.2 ml 1 M Tris-HCl (pH 7.5), 8.6 ml water

Pan-SL-RC-biotin oligo: stock at 1 nmol/µl in 10mM Tris-HCl (pH 7.5)

Oligo solution master mix was made by mixing 24 µl of 1x wash/binding buffer with 1µl of 1 nmol/ul stock per reaction. SL mRNA isolation was done using 10 µl aliquots of NEB magnetic streptavidin beads (S1420) for each reaction which were washed with 100 µl of 1x wash/binding buffer and resuspended in 25 µl of oligo solution. The streptavidin beads and biotin-tagged oligo are incubated together for 5 minutes gently shaking. They are then applied to a magnet and washed twice with 100 µl 1x wash/binding buffer. During this time, the RNA sample was prepared. The 10 µl of RNA

from an infected cow were brought up to a volume of 25 µl with water and then 25 µl of 2x wash/binding buffer added. Secondary RNA structure is dissolved by incubating the RNA sample at 65 ºC for 5 min and then placed on ice for 3 minutes to cool before proceeding with reaction. The total RNA sample was added to the prepared magnetic beads and incubated at room temperature for 2 hours gently shaking. After incubation, each sample is then washed twice with 100 µl of 1x wash/binding buffer, moved to a clean tube, and washed once more with 100 µl of ice cold 0.2M low salt buffer. All of the supernatant must be removed prior to elution and fragmentation so each tube was quickly spun down and residual wash buffer removed with a 10 µl pipette tip. SL-selected RNA was eluted from the beads by adding 15.5 µl of first strand synthesis reaction buffer and priming mixture from the NEBNext Ultra II Directional Library Preparation kit for Illumina (E7760), incubating at 94°C for 15 minutes, quickly spinning the tube, applying to a magnet, and collecting 10 µl of eluate.

Immediately proceed to use the 10 µl of enriched RNA eluted from the beads in the first strand cDNA synthesis according to manufacturer instructions. All steps subsequent steps of the library protocol should be performed as instructed in the NEBNext manual. However, since low yields were a concern for this experiment when using field isolate samples, 20 cycles of PCR were used in the library amplification step instead of the recommended 15-cycle maximum. However, this change should not be applied in future applications of this protocol. While library yields were very high, all of the prepared libraries suffered from adaptor dimer contamination which had to be removed through multiple rounds of 0.9x size selective magnetic bead clean-ups. The equimolar pooled libraries underwent six consecutive 0.9x clean-ups with Omega Mag-

Bind microbeads before sufficient adaptor dimer was removed. Initial library yields were high, so there was enough material after clean-up to submit for sequencing. Future library preparations should adhere to the 15-cycle maximum defined in the NEBNext library protocol for the PCR amplification step.

References

1. Steverding, D. (2008). The history of African trypanosomiasis. *Parasites & Vectors*, *1*(1), 3. doi:10.1186/1756-3305-1-3
2. Holmes, P. (2013). Tsetse-transmitted trypanosomes--their biology, disease impact and control. *Journal of Invertebrate Pathology*, *112 Suppl*, S11–4. doi:10.1016/j.jip.2012.07.014
3. WHO Expert Committee on the Control, Surveillance of Human African Trypanosomiasis, & World Health Organization. (2013). *Control and Surveillance of Human African Trypanosomiasis: Report of a WHO Expert Committee*(No. 984). World Health Organization.
4. Auty, H., Torr, S. J., Michoel, T., Jayaraman, S., & Morrison, L. J. (2015). Cattle trypanosomosis: the diversity of trypanosomes and implications for disease epidemiology and control. *Revue Scientifique et Technique de l'OIE*, *34*(2), 587–598. doi:10.20506/rst.34.2.2382
5. Van den Bossche, P., & Delespaux, V. (2011). Options for the control of tsetse-transmitted livestock trypanosomosis. An epidemiological perspective. *Veterinary Parasitology*, *181*(1), 37–42. doi:10.1016/j.vetpar.2011.04.021
6. Barrett, M. P., Vincent, I. M., Burchmore, R. J. S., Kazibwe, A. J. N., & Matovu, E. (2011). Drug resistance in human African trypanosomiasis. *Future Microbiology*, *6*(9), 1037–1047. doi:10.2217/fmb.11.88

7. Schwede, A., & Carrington, M. (2010). Bloodstream form Trypanosome plasma membrane proteins: antigenic variation and invariant antigens. *Parasitology*, *137*(14), 2029–2039. doi:10.1017/S0031182009992034
8. Mugnier, M. R., Stebbins, C. E., & Papavasiliou, F. N. (2016). Masters of Disguise: Antigenic Variation and the VSG Coat in Trypanosoma brucei. *PLoS Pathogens*, *12*(9), e1005784. doi:10.1371/journal.ppat.1005784
9. Mugnier, M. R., Cross, G. A. M., & Papavasiliou, F. N. (2015). The in vivo dynamics of antigenic variation in Trypanosoma brucei. *Science*, *347*(6229), 1470–1473. doi:10.1126/science.aaa4502
10. Jackson, D. G., Owen, M. J., & Voorheis, H. P. (1985). A new method for the rapid purification of both the membrane-bound and released forms of the variant surface glycoprotein from Trypanosoma brucei. *The Biochemical Journal*, *230*(1), 195–202.
11. Cross, G. A. M., Kim, H.-S., & Wickstead, B. (2014). Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427. *Molecular and Biochemical Parasitology*, *195*(1), 59–73. doi:10.1016/j.molbiopara.2014.06.004
12. Hall, J. P. J., Wang, H., & Barry, J. D. (2013). Mosaic VSGs and the scale of Trypanosoma brucei antigenic variation. *PLoS Pathogens*, *9*(7), e1003502. doi:10.1371/journal.ppat.1003502
13. Greif, G., Ponce de Leon, M., Lamolle, G., Rodriguez, M., Piñeyro, D., Tavares-Marques, L. M., … Alvarez-Valin, F. (2013). Transcriptome analysis of the

bloodstream stage from the parasite Trypanosoma vivax. *BMC Genomics*, *14*, 149. doi:10.1186/1471-2164-14-149

14. Jackson, A. P., Berry, A., Aslett, M., Allison, H. C., Burton, P., Vavrova-Anderson, J., … Berriman, M. (2012). Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species.*Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3416–3421. doi:10.1073/pnas.1117313109
15. Jackson, A. P., Goyard, S., Xia, D., Foth, B. J., Sanders, M., Wastling, J. M., … Berriman, M. (2015). Global Gene Expression Profiling through the Complete Life Cycle of Trypanosoma vivax. *PLoS Neglected Tropical Diseases*, *9*(8), e0003975. doi:10.1371/journal.pntd.0003975
16. Kolev, N. G., Ullu, E., & Tschudi, C. (2015). Construction of Trypanosoma brucei Illumina RNA-Seq libraries enriched for transcript ends. *Methods in Molecular Biology*, *1201*, 165–175. doi:10.1007/978-1-4939-1438-8_9
17. Silvester, E., Ivens, A., & Matthews, K. R. (2018). A gene expression comparison of Trypanosoma brucei and Trypanosoma congolense in the bloodstream of the mammalian host reveals species-specific adaptations to density-dependent development. *PLoS Neglected Tropical Diseases*, *12*(10), e0006863. doi:10.1371/journal.pntd.0006863
18. Camara, M., Camara, O., Ilboudo, H., Sakande, H., Kaboré, J., N'Dri, L., … Bucheton, B. (2010). Sleeping sickness diagnosis: use of buffy coats improves the sensitivity of the mini anion exchange centrifugation test.*Tropical Medicine & International Health*, *15*(7), 796–799. doi:10.1111/j.1365-3156.2010.02546.x

19. Mulindwa, J., Leiss, K., Ibberson, D., Kamanyi Marucha, K., Helbig, C., Melo do Nascimento, L., … Clayton, C. (2018). Transcriptomes of Trypanosoma brucei rhodesiense from sleeping sickness patients, rodents and culture: Effects of strain, growth conditions and RNA preparation methods. *PLoS Neglected Tropical Diseases*, *12*(2), e0006280. doi:10.1371/journal.pntd.0006280
20. Rainen, L., Oelmueller, U., Jurgensen, S., Wyrich, R., Ballas, C., Schram, J., … Tryon, V. (2002). Stabilization of mRNA expression in whole blood samples. *Clinical Chemistry*, *48*(11), 1883–1890.
21. Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., … Yaspo, M.-L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics*, *15*, 675. doi:10.1186/1471-2164-15-675
22. Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research*, *322*(1), 12–20. doi:10.1016/j.yexcr.2014.01.008
23. Günzl, A. (2010). The pre-mRNA splicing machinery of trypanosomes: complex or simplified? *Eukaryotic Cell*, *9*(8), 1159–1170. doi:10.1128/EC.00113-10
24. Liang, X., Haritan, A., Uliel, S., & Michaeli, S. (2003). trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryotic Cell*, *2*(5), 830–840. doi:10.1128/EC.2.5.830-840.2003
25. Perry, K. L., Watkins, K. P., & Agabian, N. (1987). Trypanosome mRNAs have unusual “cap 4” structures acquired by addition of a spliced leader. *Proceedings*

of the National Academy of Sciences of the United States of America, *84*(23), 8190–8194.

26. Cuypers, B., Domagalska, M. A., Meysman, P., Muylder, G. de, Vanaerschot, M., Imamura, H., … Dujardin, J.-C. (2017). Multiplexed Spliced-Leader Sequencing: A high-throughput, selective method for RNA-seq in Trypanosomatids. *Scientific Reports*, *7*(1), 3725. doi:10.1038/s41598-017-03987-0
27. González-Andrade, P., Camara, M., Ilboudo, H., Bucheton, B., Jamonneau, V., & Deborggraeve, S. (2014). Diagnosis of trypanosomatid infections: targeting the spliced leader RNA. *The Journal of Molecular Diagnostics*,*16*(4), 400–404. doi:10.1016/j.jmoldx.2014.02.006
28. Aslett et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Research 2010 38(Database issue):D457-D462; doi:10.1093/nar/gkp851
29. Storvall, H., Ramsköld, D., & Sandberg, R. (2013). Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses. *Plos One*, *8*(1), e53822. doi:10.1371/journal.pone.0053822
30. Hernández, R., & Cevallos, A. M. (2014). Ribosomal RNA gene transcription in trypanosomes. *Parasitology Research*, *113*(7), 2415–2424. doi:10.1007/s00436-014-3940-7
31. Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 11.6, Processing of rRNA and tRNA. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21729/

32. Albretsen, C., Haukanes, B. I., Aasland, R., & Kleppe, K. (1988). Optimal conditions for hybridization with oligonucleotides: a study with myc-oncogene DNA probes. *Analytical Biochemistry*, *170*(1), 193–202. doi:10.1016/0003-2697(88)90108-X

33. Tan, Z.-J., & Chen, S.-J. (2006). Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length. *Biophysical Journal*, *90*(4), 1175–1190. doi:10.1529/biophysj.105.070904

34. Blower, M. D., Jambhekar, A., Schwarz, D. S., & Toombs, J. A. (2013). Combining different mRNA capture methods to analyze the transcriptome: analysis of the Xenopus laevis transcriptome. *Plos One*, *8*(10), e77700. doi:10.1371/journal.pone.0077700

35. Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S., & Cross, G. A. M. (2010). Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Research*, *38*(15), 4946–4957. doi:10.1093/nar/gkq237

36. Lumsden, W. H. R., & Evans, D. A. (1976). *Biology of the Kinetoplastida. Vol. 1*. Academic Press Inc.(London) Ltd., 24/28 Oval Road, London NW1 7DX.

37. Wirtz E, Leal S, Ochatt C, Cross GA. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. Molecular and biochemical parasitology. 1999;99(1):89–101.

38. Gong, H., Sun, L., Chen, B., Han, Y., Pang, J., Wu, W., … Zhang, T.-M. (2016). Evaluation of candidate reference genes for RT-qPCR studies in three metabolism

related tissues of mice after caloric restriction. *Scientific Reports*, *6*, 38513. doi:10.1038/srep38513

Curriculum Vitae:

Jaime So

Date of Birth: Mar. 16, 1994

Location: Santa Rosa, California

EDUCATION

Jun. 2019	**Master of Science (Sc.M.)** **Johns Hopkins Bloomberg School of Public Health (JHSPH), Baltimore, MD** Department of Molecular Microbiology and Immunology
Jun. 2016	**Bachelor of Science (B.S.)** **University of California, Santa Barbara (UCSB)** Microbiology

RESEARCH EXPERIENCE

Oct. 2017 – present	**Research on Surface Antigen Expression of African Trypanosomes** JHSPH, Molecular Microbiology and Immunology Principal Investigator: Dr. Monica Mugnier ▪ Developed an RNA-seq library preparation protocol to enrich for kinetoplastid mRNA transcripts in RNA extracted from infected host tissues or blood.

- Analyzed the gene expression of Trypanosomes using publicly available genome references and sequencing alignment tools such as bowtie.
- Performed statistical and graphical data analysis using R programming.

Sep. 2014 – Jun. 2016 **Microbiology Research on Contact-Dependent Growth Inhibition**

UCSB, Molecular Cellular and Developmental Biology

Principal Investigators: Dr. David Low and Dr. Chris Hayes

- Participated in a project to deduce the biochemical pathways of bacterial type V secretion system toxin entry and activation within susceptible cells.
- Used Ultraviolet radiation to produce mutant strains resistant to specific toxins, then identified the gene and variant protein responsible for toxin resistance.

PEER REVIEWED PUBLICATIONS

A. M. Jones, F. Garza-Sánchez, J. So, C. S. Hayes, D. A. Low, Activation of contact-dependent antibacterial tRNase toxins by translation elongation factors. *Proceedings of the National Academy of Sciences*, 201619273 (2017).

AWARDS & HONORS

2018	**Emergent BioSolutions Fellowship**
2016	**Honors Distinction in the Major**
2015	**Undergraduate Research and Creative Activities Grant**

WORK EXPERIENCE

Oct. 2017 – present **Johns Hopkins Malaria Research Institute Insectary Core Facility**

Supervisor: Christopher Kizito

- Maintain insectary facility
- Clean cages used to rear mosquitoes for research purposes

Feb. – Jul. 2017 **Redwood Toxicology Laboratory Technician I**

Supervisor: Lister Macharia

- Urine and oral fluid drug testing. Extraction of individual samples, adding reagents, and preparing for gas and liquid chromatography.